

# Sufficient Dimension Reduction and Modeling Responses Conditioned on Covariates: An Integrated Approach via Convex Optimization

Armeen Taeb<sup>†</sup> and Venkat Chandrasekaran<sup>†,‡,\*</sup>

<sup>†</sup> Department of Electrical Engineering

<sup>‡</sup> Department of Computing and Mathematical Sciences  
California Institute of Technology  
Pasadena, Ca 91125

August 19, 2015

## Abstract

Given observations of a collection of covariates and responses  $(Y, X) \in \mathbb{R}^p \times \mathbb{R}^q$ , sufficient dimension reduction (SDR) techniques aim to identify a mapping  $f: \mathbb{R}^q \rightarrow \mathbb{R}^k$  with  $k \ll q$  such that  $Y|f(X)$  is independent of  $X$ . The image  $f(X)$  summarizes the relevant information in a potentially large number of covariates  $X$  that influence the responses  $Y$ . In many contemporary settings, the number of responses  $p$  is also quite large, in addition to a large number  $q$  of covariates. This leads to the challenge of fitting a succinctly parameterized statistical model to  $Y|f(X)$ , which is a problem that is usually not addressed in a traditional SDR framework. In this paper, we present a computationally tractable convex relaxation based estimator for simultaneously (a) identifying a linear dimension reduction  $f(X)$  of the covariates that is sufficient with respect to the responses, and (b) fitting several types of structured low-dimensional models – factor models, graphical models, latent-variable graphical models – to the conditional distribution of  $Y|f(X)$ . We analyze the consistency properties of our estimator in a high-dimensional scaling regime. We also illustrate the performance of our approach on a newsgroup dataset and on a dataset consisting of financial asset prices.

**Keywords:**  $\ell_1$  norm regularization; nuclear norm regularization; Graphical Lasso; high-dimensional inference; algebraic statistics.

## 1 Introduction

Sufficient dimension reduction (SDR) is a framework for identifying a low-dimensional approximation of a large collection of covariates that is sufficient for predicting a set of responses (Li, 1991; Duan & Li, 1991; Cook & Weisberg, 1991). Given covariates  $X \in \mathbb{R}^q$  and responses  $Y \in \mathbb{R}^p$ , the objective of SDR is to obtain a mapping  $f: \mathbb{R}^q \rightarrow \mathbb{R}^k$  with  $k \ll q$  such that the responses  $Y$  are independent of the covariates  $X$  conditioned on  $f(X)$ ; equivalently, the conditional distribution

---

\*Email: ataeb@caltech.edu, venkatc@caltech.edu

of  $Y|f(X)$  is the same as that of  $Y|X$ . The image  $f(X)$  is called a dimension reduction of the covariates  $X$  that is sufficient with respect to the responses  $Y$ , and it summarizes the relevant information in  $X$  that influences  $Y$ .

In many contemporary settings, the number of responses  $Y$  is also quite large (in addition to a potentially large number of covariates  $X$ ). For example, in financial modeling applications the responses are the prices of financial assets (numbering in the several hundreds) and the covariates may be macroeconomic indicators (see the numerical experiment in Section 3). In gene microarray analysis, the responses correspond to the expression levels of a large number of genes (on the order of tens of thousands), and the covariates could be a collection of physiological attributes (Cheung & Spielman, 2002; Brem & Kruglyak, 2005). In these problem domains, the conditional distribution of  $Y|f(X)$  is specified using a large number of parameters, and as a result the estimation of this conditional distribution leads to several statistical and computational challenges – e.g., problems with overfitting when given a modest number of observations, and difficulties with developing algorithmic procedures that operate within a reasonable computational budget. These challenges in high-dimensional inference are by now well-recognized (Bühlmann & van de Geer, 2011; Wainwright, 2014), and they have been addressed in several settings based on approximations of high-dimensional distributions consisting of many degrees of freedom by elements from structured classes of models specified using a small number of parameters; examples of such structured families include models described by sparse or banded covariance matrices (Bickel & Levina, 2008a,b; El Karoui, 2008), graphical models (Yuan & Lin, 2006; Friedman et al., 2008; Ravikumar et al., 2008; Rothman et al., 2008), and factor models (Fan et al., 2008); see Wainwright (2014) for a more extensive list of references.

## 1.1 Our Contributions

Building on this prior literature, we describe a new methodology based on convex optimization that integrates sufficient dimension reduction with techniques to fit succinctly parameterized models to the high-dimensional conditional distribution of the responses given the covariates. In particular, given observations of a set of responses  $Y$  and covariates  $X$ , we fit a linear Gaussian model with the following two properties: (a) there exists a low-dimensional linear dimension reduction<sup>1</sup>  $f(X)$  of the covariates that is sufficient with respect to the responses, and (b) the conditional distribution of  $Y|f(X)$  is specified by a concisely parameterized statistical model – the three concrete examples that we consider are a factor model, a graphical model, and a latent-variable graphical model.

If  $(Y, X) \in \mathbb{R}^p \times \mathbb{R}^q$  is a jointly Gaussian random vector with covariance matrix  $\Sigma = \begin{pmatrix} \Sigma_Y & \Sigma_{YX} \\ \Sigma'_{YX} & \Sigma_X \end{pmatrix}$ , the existence of a dimension- $k$  linear projection of the covariates  $X$  that is sufficient with respect to the responses  $Y$  is equivalent to the rank of the cross-covariance matrix  $\Sigma_{YX}$  being at most  $k$  (of interest here is the setting in which  $k < \min\{p, q\}$ ). In particular, the sufficient dimension reduction  $f(X)$  in such cases is given by the  $k$ -dimensional row-space of the matrix  $\Sigma_{YX} \cdot \Sigma_X^{-1}$ , which is the mapping that specifies the best linear estimator of  $Y$  based on  $X$ . On the other hand, the conditional statistics of the responses given the covariates are specified by the submatrix  $[\Sigma^{-1}]_Y$  of the joint precision matrix  $\Sigma^{-1}$ . As a result, the problem of fitting a concisely parameterized model to the conditional distribution of the responses given the covariates is more conveniently specified in terms of structured approximations of a submatrix of the precision

---

<sup>1</sup>If the covariates and responses are jointly Gaussian, it suffices to consider dimension reductions specified by linear mappings of the covariates (Li, 1991; Chechik et al., 2005). Even in more general settings, many approaches to SDR construct linear dimension reductions of the covariates for computational reasons (Li, 1991; Duan & Li, 1991; Cook & Weisberg, 1991; Li, 1992; Cook & Ni, 2005)

matrix  $\Sigma^{-1}$ ; for example, fitting a graphical model to the conditional distribution of the responses given the covariates corresponds to approximating the submatrix  $[\Sigma^{-1}]_Y$  by a sparse matrix. The different parameterizations in which these two modeling tasks are most naturally described – SDR in terms of covariance matrices and conditional modeling of the responses given the covariates in terms of precision matrices – poses an obstruction to their integration into a single framework.

We overcome this difficulty by making the observation that  $\text{rank}(\Sigma_{YX}) = \text{rank}([\Sigma^{-1}]_{YX})$  in non-degenerate models (i.e., the joint covariance matrix  $\Sigma$  is positive definite) based on the following relation between the two alternative forms – in terms of the precision matrix and in terms of the covariance matrix – of the mapping that specified the best linear estimate of  $Y$  based on  $X$ :

$$\Sigma_{YX} \cdot \Sigma_X^{-1} = -[\Sigma^{-1}]_Y \cdot [\Sigma^{-1}]_{YX} \quad (1.1)$$

Hence, our approach for integrating SDR and conditional modeling of the responses given the covariates is to fit a Gaussian model specified by a precision matrix  $\Theta = \begin{pmatrix} \Theta_Y & \Theta_{YX} \\ \Theta'_{YX} & \Theta_X \end{pmatrix}$  to a collection of observations of covariates and responses such that (a) the submatrix  $\Theta_{YX}$  is low-rank, and (b) the submatrix  $\Theta_Y$  (which specifies the conditional distribution of the responses given the covariates) has a concise parameterization that describes a structured model. This reformulation of our modeling framework in terms of precision matrices leads naturally to the following general form of an estimator, given joint observations  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n \subset \mathbb{R}^{p+q}$  of the responses and covariates:

$$\underset{\Theta \in \mathbb{S}^{p+q}, \Theta \succ 0}{\text{argmin}} \quad -\ell(\Theta; \{Y^{(i)}, X^{(i)}\}_{i=1}^n) + \lambda_n[\gamma \|\Theta_{YX}\|_* + R(\Theta_Y)] \quad (1.2)$$

The set  $\mathbb{S}^k$  denotes the space of  $k \times k$  symmetric matrices, the function  $\ell(\Theta; \{X^{(i)}, Y^{(i)}\}_{i=1}^n)$  denotes the log-likelihood of the observations  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n$  with respect to a Gaussian distribution parameterized by the precision matrix  $\Theta$ , the function  $\|\cdot\|_*$  denotes the nuclear norm (sum of the singular values of a matrix), and the function  $R : \mathbb{S}^p \rightarrow \mathbb{R}$  is suitably chosen to promote a desired structure in the submatrix  $\Theta_Y$  (i.e., the conditional distribution of the responses given the covariates). The role of the nuclear norm penalty is to promote low rank structure in the submatrix  $\Theta_{YX}$ ; this regularizer has been successfully employed in many settings for fitting structured low-rank models to high-dimensional data (Fazel, 2002; Recht et al., 2009). Here  $\lambda_n > 0$  and  $\gamma > 0$  are regularization parameters. By virtue of the convexity of norms and of the negative of the log-likelihood function  $\ell(\Theta; \{Y^{(i)}, X^{(i)}\}_{i=1}^n)$ , the program (1.2) is a convex optimization problem if the function  $R(\Theta_Y)$  is convex. In the following discussion, we provide three concrete approaches to fit structured low-dimensional models to the conditional distribution of  $Y|f(X)$  via suitable choices of the function  $R(\Theta_Y)$ .

**SDR + Factor Modeling (SDR-FM)** Fitting a factor model to the conditional distribution of  $Y|f(X)$  corresponds to approximating the covariance matrix of  $Y|f(X)$  as the sum of a diagonal matrix and a low-rank matrix. By appealing to the Sherman-Morrison-Woodbury formula (Horn & Johnson, 1990), the covariance matrix of  $Y|f(X)$  being decomposable as the sum of a diagonal matrix and a low-rank matrix is equivalent to the precision matrix of  $Y|f(X)$  being decomposable as the *difference* between a diagonal matrix and a low-rank matrix. Thus, a natural choice for the regularizer  $R(\Theta_Y)$  is:

$$R(\Theta_Y) = \inf_{D_Y, L_Y \in \mathbb{S}^p} \text{trace}(L_Y) \quad \text{subject to} \quad \Theta_Y = D_Y - L_Y, \quad L_Y \succeq 0, \quad D_Y \text{ is diagonal.} \quad (1.3)$$

Here  $D_Y, \Theta_Y$  represent the diagonal and low-rank components of  $\Theta_Y$ . As before, the role of the nuclear norm penalty (the nuclear norm for positive semidefinite matrices reduces to the trace) is

to enforce low-rank structure in the  $L_Y$  component.

**SDR + Graphical Modeling (SDR-GM)** Fitting a sparse graphical model to the conditional distribution of  $Y|f(X)$  corresponds to approximating the submatrix  $\Theta_Y$  by a sparse matrix; the sparsity pattern of  $\Theta_Y$  specifies the graphical model structure underlying the conditional distribution of  $Y|f(X)$ . Based on prior work on the Graphical Lasso (Yuan & Lin, 2006; Friedman et al., 2008), an appropriate choice for the regularizer  $R(\Theta_Y)$  is:

$$R(\Theta_Y) = \|\Theta_Y\|_{\ell_1}. \quad (1.4)$$

The function  $\|\cdot\|_{\ell_1}$  denotes the  $\ell_1$  norm (sum of the magnitudes of the entries of a matrix), and its role is to induce sparsity in the submatrix  $\Theta_Y$ . Regularizers based on the  $\ell_1$  norm have been widely and successfully employed in many settings for fitting sparse models to data in high dimensions (Tibshirani, 1996; Chen et al., 2001; Candès et al., 2006; Donoho, 2006). The convex program (1.2) with  $R(\Theta_Y) = \|\Theta_Y\|_{\ell_1}$  is a natural extension of the Graphical Lasso (Yuan & Lin, 2006; Friedman et al., 2008) in which an  $\ell_1$  norm penalty is employed to induce sparsity in the precision matrix (although the Graphical Lasso operates purely on observations of a set of responses and it does not consist of a nuclear norm penalty corresponding to an SDR objective).

**SDR + Latent-Variable Graphical Modeling (SDR-LVGM)** We also describe a generalization of the SDR-GM approach. In practice, it may be expensive or infeasible for a data analyst to gather observations of all the relevant covariates that may potentially impact the responses. As a result, the responses  $Y$  could be affected by unobserved latent variables in addition to being influenced by the covariates  $X$ . These latent variables can lead to confounding dependencies among the responses, which in turn complicates the task of fitting a graphical model to the conditional distribution of  $Y|f(X)$ . Motivated by these considerations, it is of interest to fit a graphical model to the distribution of  $Y|f(X), \zeta$ , where  $\zeta \in \mathbb{R}^h$  (here  $h \ll p$ ) represent a small number of latent variables that are statistically independent of  $f(X)$ . As described by Chandrasekaran et al. (2012), fitting such a latent-variable graphical model corresponds to approximating the submatrix  $\Theta_Y$  by the sum of a sparse matrix and a low-rank matrix, rather than just a sparse matrix as in the case of a pure graphical model approximation. Here the low-rank component of  $\Theta_Y$  accounts for the effect of the latent variables  $\zeta \in \mathbb{R}^h$  on the responses  $Y$ , and the rank of this component is equal to the dimension  $h$  of  $\zeta$ . The sparse component of  $\Theta_Y$  specifies the graphical model structure underlying  $Y|f(X), \zeta$ . Building on the insights in Chandrasekaran et al. (2012), the regularizer  $R_\delta(\Theta_Y)$  in this setting chosen as:

$$R_\delta(\Theta_Y) = \inf_{S_Y, L_Y \in \mathbb{S}^p} \text{trace}(L_Y) + \delta \|S_Y\|_{\ell_1} \quad \text{subject to} \quad \Theta_Y = S_Y - L_Y, \quad L_Y \succeq 0. \quad (1.5)$$

The matrices  $S_Y, L_Y \in \mathbb{S}^p$  in (1.5) correspond to the sparse and low-rank components of  $\Theta_Y$ , respectively, and  $\delta > 0$  is a regularization parameter. As before, the  $\ell_1$  norm and the nuclear norm penalties promote the type of structure that we desire in our model. Plugging in the regularizer  $R_\delta(\Theta_Y)$  into (1.2), we obtain the following optimization program for joint SDR and latent-variable graphical modeling given observations  $\{Y^{(i)}, X^{(i)}\}$  of the responses and the covariates:

$$\begin{aligned} (\hat{\Theta}, \hat{S}_Y, \hat{L}_Y) = & \underset{\substack{\Theta \in \mathbb{S}^{p+q}, \Theta \succ 0 \\ S_Y, L_Y \in \mathbb{S}^p}}{\text{argmin}} & -\ell(\Theta; \{Y^{(i)}, X^{(i)}\}_{i=1}^n) + \lambda_n [\gamma \|\Theta_{YX}\|_* + \text{trace}(L_Y) + \delta \|S_Y\|_{\ell_1}] \\ \text{s.t.} & & \Theta_Y = S_Y - L_Y, \quad L_Y \succeq 0. \end{aligned} \quad (1.6)$$

The convex program (1.6) is a natural extension of the estimator proposed by Chandrasekaran et al. (2012) in which a trace penalty and an  $\ell_1$  norm penalty are employed to fit a latent-variable

graphical model to a set of responses. To be clear, the estimator (1.6) also incorporates SDR to identify a low-dimensional projection of a set of covariates that are sufficient for predicting the responses, which is in contrast to the estimator proposed by Chandrasekaran et al. (2012).

The regularizers defined in (1.3), (1.4), and (1.5) are convex with respect to the submatrix  $\Theta_Y$ . As a result, the estimators corresponding to the SDR-FM, SDR-GM, and SDR-LVGM approaches are convex optimization programs. The estimator (1.6) corresponding to the SDR-LVGM approach is in some sense a generalization of the estimators corresponding to the SDR-GM and SDR-FM approaches, as latent-variable graphical modeling may be viewed as a blend of factor modeling and graphical modeling. As a result, we focus in Section 2 on analyzing the consistency properties of the estimator (1.6) in a high-dimensional scaling regime. Specifically, suppose we observe samples  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n \subset \mathbb{R}^{p+q}$  of a collection of jointly Gaussian responses and covariates  $(Y, X) \in \mathbb{R}^{p+q}$ ,

with population precision matrix  $\Theta^* = \begin{pmatrix} S_Y^* - L_Y^* & \Theta_{YX}^* \\ \Theta_{YX}^{*\prime} & \Theta_X^* \end{pmatrix} \in \mathbb{S}^{p+q}$ . Supplying these observations

as input to the convex program (1.6) and obtaining estimates  $(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y) \in \mathbb{S}^{p+q} \times \mathbb{S}^p \times \mathbb{S}^p$ , we prove in Theorem 2.1 that (under certain conditions on  $\Theta^*$  and with high probability) the rank of the submatrix  $\hat{\Theta}_{YX}$  is equal to the rank of  $\Theta_{YX}^*$ , the rank of  $\hat{L}_Y$  is equal to the rank of  $L_Y^*$ , and the sparsity pattern of  $\hat{S}_Y$  is the same as that of  $S_Y^*$ ; these recovery guarantees imply that we obtain the correct dimension of the image  $f(X)$  specifying the sufficient dimension reduction of the covariates, the correct number of unobserved latent variables, and the correct conditional graphical model structure underlying the population. Informally, the assumptions on the population precision matrix  $\Theta^*$  are that: (a) the submatrix  $\Theta_{YX}^*$  is sufficiently low-rank; (b) the submatrix  $\Theta_Y^* = S_Y^* - L_Y^*$  is such that  $S_Y^*$  is sufficiently sparse and  $L_Y^*$  is sufficiently low-rank; and (c) the population Fisher information  $\Theta^{*-1} \otimes \Theta^{*-1}$  obeys certain irrepresentability-type conditions; see Assumptions 1 and 2 in Section 2.2, Theorem 2.1 and the subsequent discussion in Appendix 5.2 for a precise formulation of these conditions. The first assumption above on  $\Theta^*$  states that there exists a low-dimensional linear projection  $f(X)$  of the covariates  $X$  that is sufficient with respect to  $Y$ . The second condition requires that  $Y|f(X)$  is specified by a latent-variable graphical model with a small number of latent variables and a sparse graphical model. The third assumption is analogous to the irrepresentability conditions that play a role in the analysis of the consistency of the Lasso (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Wainwright, 2014), the Graphical Lasso (Ravikumar et al., 2008), the convex relaxation proposed by Chandrasekaran et al. (2012) for latent-variable graphical modeling, as well as other estimators in high-dimensional inference problems (Bühlmann & van de Geer, 2011; Wainwright, 2014).

In Section 3 we illustrate the performance of the estimators corresponding to the SDR-FM, SDR-GM, and SDR-LVGM approaches on two datasets. First, we consider a financial asset modeling problem in which the responses are a collection of stock returns of 67 companies from the Standard and Poor index, and the covariates are the following 7 macroeconomic indicators: the industrial production index, the inflation rate, the amount of oil exports, the population growth rate, the unemployment rate, the consumer credit score, and the EUR to USD exchange rate. In the second experiment, we analyze the 20newsgroup dataset that consists of 16,242 samples in  $\mathbb{R}^{100}$ , with each observation corresponding to a news document. The coordinates of these observations are indexed by a collection of 100 words, and each observation is a binary vector specifying whether a word appears in the document. Of those 100 words, the following 9 words are chosen to be the covariates as they appear to be useful in categorizing newsgroup documents: government, religion, science, technology, war, medicine, world, food, and games. The remaining 91 words are used as response variables.

**Adapting to Alternative Forms of SDR** In many settings, one is interested in identifying a *subset* of the covariates  $X$  that is useful for predicting the responses  $Y$ , rather than a generic dimension reduction of  $X$ . In such cases, it is more natural to fit a linear Gaussian model with a precision matrix  $\Theta \in \mathbb{S}^{p+q}$  in which the submatrix  $\Theta_{YX}$  is column-sparse (i.e, only a subset of the columns of this matrix are nonzero) instead of being low-rank. This point follows from the relation (1.1) by noting that the map  $-\lceil \Sigma^{-1} \rceil_Y \cdot \lceil \Sigma^{-1} \rceil_{YX}$  specifying the best linear estimator of  $Y$  based on  $X$  must be column-sparse if a subset of the covariates is sufficient for predicting the responses (the indices of the nonzero columns of  $\Theta_{YX}$  correspond to the subset of the covariates that are relevant for predicting the responses). To fit a model with a column-sparse submatrix  $\Theta_{YX}$ , one can modify the family of estimators (1.2) by replacing the nuclear norm penalty  $\|\Theta_{YX}\|_*$  with a group norm penalty  $\|\Theta_{YX}\|_{2,1} = \sum_{i=1}^q \|(\Theta_{YX})_{:,i}\|_{\ell_2}$ , where  $(\Theta_{YX})_{:,i}$  represents the  $i$ 'th column of  $\Theta_{YX}$ . Such group norm penalties are useful for inducing sparsity in entire columns of  $\Theta_{YX}$  (Yuan & Lin, 2006). As an illustration, given observations  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n \subset \mathbb{R}^{p+q}$  the estimator (1.6) can be modified as follows:

$$\begin{aligned}
(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y) = & \underset{\substack{\Theta \in \mathbb{S}^{p+q}, \Theta \succ 0 \\ S_Y, L_Y \in \mathbb{S}^p}}{\operatorname{argmin}} & -\ell(\Theta; \{Y^{(i)}, X^{(i)}\}_{i=1}^n) + \lambda_n[\gamma \|\Theta_{YX}\|_{2,1} + \operatorname{trace}(L_Y) + \delta \|S_Y\|_{\ell_1}] \\
& \text{s.t.} & \Theta_Y = S_Y - L_Y, \quad L_Y \succeq 0.
\end{aligned} \tag{1.7}$$

This estimator simultaneously identifies a subset of the covariates that are relevant for predicting the responses and also fits a latent-variable graphical model to the conditional distribution of the responses given the covariates. The analysis of the statistical consistency of this estimator is similar in spirit to that of the estimator (1.6); see Appendix 5.3 for more details.

## 1.2 Related Work

Many researchers have developed techniques for computing sufficient dimension reductions (see the survey (Adraghi & Cook, 2009) and the references therein), with Sliced Inverse Regression being a prominent example (Li, 1991). In a jointly Gaussian setting, classical approaches such as Canonical Correlations Analysis (CCA) or Partial Least Squares (PLS) may also be employed to compute linear sufficient dimension reductions (Fung et al., 2002), although the objectives of CCA and PLS are somewhat different than that of SDR. More recently, Negahban & Wainwright (2011) employed a nuclear norm penalty in a multivariate linear regression setup to identify a low-dimensional projection of a set of covariates that best predicts a set of responses. However, none of these papers consider the additional challenge of modeling the conditional distribution of the responses given the covariates.

A number of researchers have developed methods for simultaneously obtaining a concise model of the predictive relationship of the covariates on the responses, while also fitting a sparse graphical model to the conditional distribution of the responses given the covariates. For example, the techniques introduced by Rothman et al. (2008), Cai et al. (2010), and Yin & Li (2011) may be interpreted as seeking a sparse approximation to the matrix  $\Sigma_{YX} \cdot \Sigma_X^{-1}$  (or equivalently,  $-\Theta_Y^{-1} \cdot \Theta_{YX}$ ) that specifies the best linear estimator of  $Y$  based on  $X$ , in addition to computing a sparse approximation to the submatrix  $\Theta_Y$ . The algorithms proposed in these papers are either non-convex or involve multi-step procedures consisting of several convex programs. In a different direction, Sohn & Kim (2012) and Yuan & Zhang (2014) consider a regularized log-likelihood convex program with  $\ell_1$  norm penalties on the submatrices  $\Theta_Y$  and  $\Theta_{YX}$  of the joint precision matrix. In contrast to these papers, our approach for modeling the predictive relationship of the covariates on the responses is based on SDR, where we seek a low-dimensional projection of the covariates that is sufficient for

predicting the responses. Further, our framework integrates SDR and conditional modeling of the responses given the covariates via a single convex program.

In the same work referenced above, Yuan & Zhang (2014) also consider the problem of selecting a subset of a collection of covariates that are most relevant for predicting a set of responses, while additionally fitting a sparse graphical model to the conditional distribution of the responses given the selected covariates. They address this problem by proposing the following regularized log-likelihood convex program consisting of an  $\|\cdot\|_{2,1}$  norm penalty on  $\Theta_{YX}$  and an  $\ell_1$  norm penalty on  $\Theta_Y$ :

$$\hat{\Theta} = \underset{\Theta \in \mathbb{S}^{p+q}, \Theta \succ 0}{\operatorname{argmin}} -\ell(\Theta; \{Y^{(i)}, X^{(i)}\}_{i=1}^n) + \lambda_n[\gamma\|\Theta_{YX}\|_{2,1} + \|\Theta_Y\|_{\ell_1}] \quad (1.8)$$

In Appendix 5.3, we analyze the high-dimensional consistency of the estimator (1.7) – which fits a latent-variable graphical model to the conditional distribution of the responses given the selected covariates rather than just a graphical model – and this analysis can be specialized to obtain consistency results for the estimator (1.8). More broadly, one of the key distinctions between our work and that of Yuan & Zhang (2014) is that our approach (1.2) for simultaneous SDR and conditional modeling of responses given covariates is useful for obtaining general linear sufficient dimension reductions of the covariates rather than just selecting subsets of relevant covariates. Moreover, our framework can be adapted to fit three types of models to the conditional distribution of the responses conditioned on the covariates.

### 1.3 Notation

Given a matrix  $U \in \mathbb{R}^{p_1 \times p_2}$ , the norm  $\|U\|_{\ell_\infty}$  denotes the largest entry in magnitude of  $U$ , and the norm  $\|U\|_2$  denotes the spectral norm (the largest singular value of  $U$ ). The norm  $\|U\|_{2,\infty}$  denotes the maximum of the  $\ell_2$  norms of the columns of  $U$ :  $\|U\|_{2,\infty} = \max_{i=1,2,\dots,p_2} \|U_{:,i}\|_{\ell_2}$ . We denote the set of  $k \times k$  positive semidefinite and positive-definite matrices by  $\mathbb{S}_+^k$  and  $\mathbb{S}_{++}^k$ , respectively. Finally, the linear operators  $\mathcal{A} : \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q \rightarrow \mathbb{S}^{(p+q)}$  and its adjoint  $\mathcal{A}^\dagger : \mathbb{S}^{(p+q)} \rightarrow \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$  are defined as follows:

$$\mathcal{A}(M, N, K, O) \triangleq \begin{pmatrix} M - N & K \\ K^T & O \end{pmatrix}, \quad \mathcal{A}^\dagger \begin{pmatrix} Q & K \\ K^T & O \end{pmatrix} \triangleq (Q, Q, K, O) \quad (1.9)$$

## 2 Model Selection Consistency

As described in the introduction, we investigate the consistency properties of the SDR-LVGM estimator (1.6), which integrates SDR and latent-variable graphical modeling of the conditional distribution of the responses given the covariates. The main result is stated in Section 2.3, and it is based on assumptions on the population precision matrix (discussed in Section 2.1) and irrepresentability-type conditions on the population Fisher information (discussed in Section 2.2).

### 2.1 Technical Setup

Let  $\Theta^* = \begin{pmatrix} \Theta_Y^* & \Theta_{YX}^* \\ \Theta_{YX}^{*\top} & \Theta_X^* \end{pmatrix} \in \mathbb{S}_{++}^{p+q}$  denote the precision matrix of a jointly Gaussian random vector  $(Y, X) \in \mathbb{R}^{p+q}$  of covariates and responses, and let  $\Sigma^* = \Theta^{*-1}$  denote the corresponding covariance matrix. The submatrix  $\Theta_{YX}^* \in \mathbb{R}^{p \times q}$  is a rank- $k$  matrix, where  $k \ll \min\{p, q\}$  is the size of the smallest dimension reduction  $f(X)$  of the covariates  $X$  that is sufficient with respect to the

responses  $Y$ . The submatrix  $\Theta_Y^* \in \mathbb{S}^p$  is the precision matrix of the conditional distribution of  $Y|f(X)$ , and it specifies a latent-variable graphical model (Chandrasekaran et al., 2012). That is, the matrix  $\Theta_Y^*$  is decomposed as  $\Theta_Y^* = S_Y^* - L_Y^*$  – the component  $S_Y^*$  is a sparse matrix representing the precision matrix of the distribution of  $Y$  conditioned on  $f(X)$  as well as a small number of additional unobserved latent variables  $\zeta \in \mathbb{R}^h$  (here  $h \ll p$ ), and the component  $L_Y^*$  is a low-rank matrix representing the effect of the latent variables  $\zeta$  ( $\text{rank}(L_Y^*) = h$ ). These structural attributes of our model lead to the following definition:

**Definition 2.1.** An estimate  $(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y) \in \mathbb{S}_{++}^{q+p} \times \mathbb{S}^p \times \mathbb{S}_+^p$  with  $\hat{\Theta}_Y = \hat{S}_Y - \hat{L}_Y$  is a structurally correct estimate of the model specified by the matrices  $(\Theta^*, S_Y^*, L_Y^*) \in \mathbb{S}_{++}^{q+p} \times \mathbb{S}^p \times \mathbb{S}_+^p$  with  $\Theta_Y^* = S_Y^* - L_Y^*$  if (1)  $\text{rank}(\hat{\Theta}_{YX}) = \text{rank}(\Theta_{YX}^*)$  and (2)  $\text{sign}(\hat{S}_Y) = \text{sign}(S_Y^*)$  (here  $\text{sign}(0) = 0$ );  $\text{rank}(\hat{L}_Y) = \text{rank}(L_Y^*)$ .

Condition (1) ensures that the size of the smallest dimension reduction  $f(X)$  of the covariates  $X$  that is sufficient with respect to the responses  $Y$  is estimated correctly. Condition (2) ensures that the latent-variable graphical model specifying the conditional distribution of  $Y|f(X)$  is estimated correctly, which corresponds to accurately identifying the two components composing the precision matrix  $\Theta_Y^*$  of  $Y|f(X)$ . In particular, this condition ensures that (a)  $\hat{S}_Y$  provides a structurally correct estimate of the graphical model specifying the conditional distribution of  $Y|f(X), \zeta$ ; that is, positive, negative and zero entries in  $S_Y^*$  are estimated correctly as positive, negative, and zero entries, respectively, in  $\hat{S}_Y$ , and (b) the dimension of the latent variables  $\zeta$  is estimated correctly.

Following the literature on high-dimensional estimation (see the surveys (Bühlmann & van de Geer, 2011; Wainwright, 2014) and the references therein), a natural set of conditions for obtaining consistent and structurally correct parameter estimates is to assume that the curvature of the likelihood function at  $\Theta^*$  is bounded in certain directions. This curvature is governed by the Fisher information at  $\Theta^*$ :

$$\mathbb{I}(\Theta^*) \triangleq \Theta^{*-1} \otimes \Theta^{*-1} = \Sigma^* \otimes \Sigma^*.$$

Here  $\otimes$  denotes a tensor product between matrices and  $\mathbb{I}(\Theta^*)$  may be viewed as a map from  $\mathbb{S}^{(p+q)}$  to  $\mathbb{S}^{(p+q)}$ . We impose conditions requiring that  $\mathbb{I}(\Theta^*)$  is well-behaved when applied to matrices of the form  $\Theta - \Theta^* = \begin{pmatrix} (S_Y - S_Y^*) - (L_Y - L_Y^*) & \Theta_{YX} - \Theta_{YX}^* \\ \Theta_{YX}' - \Theta_{YX}^{*'} & \Theta_X - \Theta_X^* \end{pmatrix}$ , where  $S_Y$  is in a neighborhood of  $S_Y^*$  restricted to the set of sparse matrices and  $(L_Y, \Theta_{YX})$  are in a neighborhood of  $(L_Y^*, \Theta_{YX}^*)$  restricted to sets of low-rank matrices. As formally described in Section 2.2, these local properties of  $\mathbb{I}(\Theta^*)$  around  $\Theta^*$  are conveniently stated in terms of *tangent spaces* to the algebraic varieties of sparse matrices and of low-rank matrices.

Let  $M \in \mathbb{S}^p$  be a symmetric matrix with  $k$  nonzero entries. The tangent space at  $M$  with respect to the algebraic variety of matrices in  $\mathbb{S}^p$  with at most  $k$  nonzero entries is given by:

$$\Omega(M) \triangleq \{J \in \mathbb{S}^p \mid \text{support}(J) \subseteq \text{support}(M)\}.$$

Here ‘support’ denotes the set of locations of the nonzero entries. Next, consider a rank- $r$  matrix  $N \in \mathbb{R}^{p_1 \times p_2}$  with reduced singular value decomposition (SVD) given by  $N = UDV'$ , where  $U \in \mathbb{R}^{p_1 \times r}$ ,  $D \in \mathbb{R}^{r \times r}$ , and  $V \in \mathbb{R}^{p_2 \times r}$ . The tangent space at  $N$  with respect to the algebraic variety of  $p_1 \times p_2$  matrices with rank less than or equal to  $r$  is given by<sup>2</sup>:

$$T(N) \triangleq \{UY_1' + Y_2V' \mid Y_1 \in \mathbb{R}^{p_2 \times r}, Y_2 \in \mathbb{R}^{p_1 \times r}\}.$$

<sup>2</sup>We also consider the tangent space at a symmetric low-rank matrix with respect to the algebraic variety of symmetric low-rank matrices. We use the same notation ‘ $T$ ’ to denote tangent spaces in both the symmetric and non-symmetric cases, and the appropriate tangent space is clear from the context.



In the next section, we describe irrepresentability conditions on the population Fisher information  $\mathbb{I}(\Theta^*)$  in terms of the tangent spaces  $\Omega(S_Y^*)$ ,  $T(L_Y^*)$ , and  $T(\Theta_{YX}^*)$ ; under these conditions, we prove in Appendix 5.1 that the regularized maximum-likelihood convex program (1.6) provides structurally correct and consistent estimates.

## 2.2 Fisher Information Conditions For Consistency

Let  $\mathbb{I}^* = \mathbb{I}(\Theta^*)$  denote the population Fisher information at  $\Theta^*$ . Given a norm  $\|\cdot\|_{\Upsilon}$  on  $\mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ , the first condition we consider is to bound the minimum gain of  $\mathbb{I}^*$  restricted to a subspace  $\mathbb{H} \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$  as follows:

$$\chi(\mathbb{H}, \|\cdot\|_{\Upsilon}) \triangleq \min_{\substack{(S_Y, L_Y, \Theta_{YX}, \Theta_X) \in \mathbb{H} \\ \|(S_Y, L_Y, \Theta_{YX}, \Theta_X)\|_{\Upsilon} = 1}} \|\mathcal{P}_{\mathbb{H}} \mathcal{A}^{\dagger} \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}}(S_Y, L_Y, \Theta_{YX}, \Theta_X)\|_{\Upsilon}, \quad (2.1)$$

where  $\mathcal{P}_{\mathbb{H}}$  denotes the projection operator onto the subspace  $\mathbb{H}$  and the linear maps  $\mathcal{A}$  and  $\mathcal{A}^{\dagger}$  are defined in (1.9). The quantity  $\chi(\mathbb{H}, \|\cdot\|_{\Upsilon})$  being large ensures that the Fisher information  $\mathbb{I}^*$  is well-conditioned restricted to image  $\mathcal{A}\mathbb{H} \subseteq \mathbb{S}^{p+q}$ . The second condition that we impose on  $\mathbb{I}^*$  is in the spirit of irrepresentability-type conditions (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Wainwright, 2009; Ravikumar et al., 2008; Chandrasekaran et al., 2012) that are frequently employed in high-dimensional estimation. Specifically, we require that the inner-product between elements in  $\mathcal{A}\mathbb{H}$  and  $\mathcal{A}\mathbb{H}^{\perp}$ , as quantified by the metric induced by  $\mathbb{I}^*$ , is bounded above:

$$\varphi(\mathbb{H}, \|\cdot\|_{\Upsilon}) \triangleq \max_{\substack{Z \in \mathbb{H} \\ \|Z\|_{\Upsilon} = 1}} \|\mathcal{P}_{\mathbb{H}^{\perp}} \mathcal{A}^{\dagger} \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}} (\mathcal{P}_{\mathbb{H}} \mathcal{A}^{\dagger} \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}})^{-1}(Z)\|_{\Upsilon}. \quad (2.2)$$

The operator  $(\mathcal{P}_{\mathbb{H}} \mathcal{A}^{\dagger} \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}})^{-1}$  in (2.2) is well-defined if  $\chi(\mathbb{H}, \|\cdot\|_{\Upsilon}) > 0$ , since this latter condition implies that  $\mathbb{I}^*$  is injective restricted to  $\mathcal{A}\mathbb{H}$ . The quantity  $\varphi(\mathbb{H}, \|\cdot\|_{\Upsilon})$  being small implies that any element of  $\mathcal{A}\mathbb{H}$  and any element of  $\mathcal{A}\mathbb{H}^{\perp}$  have a small inner-product (in the metric induced by  $\mathbb{I}^*$ ).

A natural approach to controlling the conditioning of the Fisher information around  $\Theta^*$  is to bound the quantities  $\chi(\mathbb{H}^*, \|\cdot\|_{\Upsilon})$  and  $\varphi(\mathbb{H}, \|\cdot\|_{\Upsilon})$  for  $\mathbb{H}^* = \Omega(S_Y^*) \times T(L_Y^*) \times T(\Theta_{YX}^*) \times \mathbb{S}^q$ . However, a complication that arises with this approach is that the varieties of low-rank matrices are locally curved around  $L_Y^*$  and around  $\Theta_{YX}^*$ . Consequently, the tangent spaces at points in neighborhoods around  $L_Y^*$  and around  $\Theta_{YX}^*$  are not the same as  $T(L_Y^*)$  and  $T(\Theta_{YX}^*)$ . (A similar difficulty does not arise with sparse matrices, as the variety of sparse matrices is locally flat around  $S_Y^*$ ; hence, the tangent spaces at all points in a neighborhood of  $S_Y^*$  are the same.) In order to account for this curvature underlying the varieties of low-rank matrices, we bound the distance between nearby tangent spaces via the following induced norm:

$$\rho(T_1, T_2) \triangleq \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T_1} - \mathcal{P}_{T_2})(N)\|_2.$$

Using this approach for bounding nearby tangent spaces, we consider subspaces  $\mathbb{H}' = \Omega(S_Y^*) \times T_Y' \times T_{YX}' \times \mathbb{S}^q$  for all  $T_Y'$  close to  $T(L_Y^*)$  and for all  $T_{YX}'$  close to  $T(\Theta_{YX}^*)$ , as measured by  $\rho$  (Chandrasekaran et al., 2012). For  $\omega_Y, \omega_{YX} \in (0, 1)$ , we bound  $\chi(\mathbb{H}', \|\cdot\|_{\Upsilon})$  and  $\varphi(\mathbb{H}', \|\cdot\|_{\Upsilon})$  in the sequel for all subspaces  $\mathbb{H}'$  in the following set:

$$U(\omega_Y, \omega_{YX}) \triangleq \left\{ \Omega(S_Y^*) \times T_Y' \times T_{YX}' \times \mathbb{S}^q \mid \begin{aligned} &\rho(T_Y', T(L_Y^*)) \leq \omega_Y \\ &\rho(T_{YX}', T(\Theta_{YX}^*)) \leq \omega_{YX} \end{aligned} \right\}. \quad (2.3)$$

We control the quantities  $\chi(\mathbb{H}', \|\cdot\|_Y)$  and  $\varphi(\mathbb{H}', \|\cdot\|_Y)$  using a slight variant of the dual norm of the regularizer  $\delta\|S_Y\|_{\ell_1} + \text{trace}(L_Y) + \gamma\|\Theta_{YX}\|_*$  in (1.6):

$$\Phi_{\delta,\gamma}(S_Y, L_Y, \Theta_{YX}, \Theta_X) \triangleq \max \left\{ \frac{\|S_Y\|_{\ell_\infty}}{\delta}, \|L_Y\|_2, \frac{\|\Theta_{YX}\|_2}{\gamma}, \|\Theta_X\|_2 \right\}. \quad (2.4)$$

As the dual norm  $\max \left\{ \frac{\|S_Y\|_{\ell_\infty}}{\delta}, \|L_Y\|_2, \frac{\|\Theta_{YX}\|_2}{\gamma} \right\}$  of the regularizer in (1.6) plays a central role in the optimality conditions of (1.6), controlling the quantities  $\chi(\mathbb{H}', \Phi_{\delta,\gamma})$  and  $\varphi(\mathbb{H}', \Phi_{\delta,\gamma})$  leads to a natural set of conditions that guarantee the structural correctness and consistency of the estimates produced by (1.6). In summary, given a fixed set of parameters  $(\delta, \gamma, \omega_Y, \omega_{YX}) \in \mathbb{R}_+ \times \mathbb{R}_+ \times (0, 1) \times (0, 1)$ , we assume that  $\mathbb{I}^*$  satisfies the following conditions:

$$\text{Assumption 1} : \inf_{\mathbb{H}' \in U(\omega_Y, \omega_{YX})} \chi(\mathbb{H}', \Phi_{\delta,\gamma}) \geq \alpha, \quad \text{for some } \alpha > 0 \quad (2.5)$$

$$\text{Assumption 2} : \sup_{\mathbb{H}' \in U(\omega_Y, \omega_{YX})} \varphi(\mathbb{H}', \Phi_{\delta,\gamma}) \leq 1 - \nu \quad \text{for some } \nu \in (0, 1/3). \quad (2.6)$$

For fixed  $(\delta, \gamma, \omega_Y, \omega_{YX})$ , larger values of  $\alpha$  and  $\nu$  in these assumptions lead to a better conditioned  $\mathbb{I}^*$ .

Assumptions 1 and 2 are analogous to conditions that play an important role in the analysis of the Lasso for sparse linear regression (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Wainwright, 2009), graphical model selection via the Graphical Lasso (Ravikumar et al., 2008), and in several other approaches for high-dimensional estimation (Bühlmann & van de Geer, 2011; Wainwright, 2014). As a point of comparison with respect to analyses of the Lasso, the role of the Fisher information  $\mathbb{I}^*$  in (2.5) and in (2.6) is played by  $A^T A$ , where  $A$  is the underlying design matrix (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Wainwright, 2009). In analyses of both the Lasso and the Graphical Lasso in the papers referenced above, the analog of the subspace  $\mathbb{H}$  is the set of models with support contained inside the support of the underlying sparse population model. Assumptions 1 and 2 are also similar in spirit to conditions employed in the analysis of convex relaxation methods for latent-variable graphical model selection (Chandrasekaran et al., 2012). However, as emphasized previously in the introduction, an important distinction between the present paper and prior literature on graphical model selection is that the methods and results in previous work are not directly applicable to the problem of simultaneous SDR and (latent-variable) graphical modeling.

## 2.3 High-Dimensional Consistency Result

In this section, we describe the performance of the regularized maximum-likelihood program (1.6). Before formally stating our main result, we introduce some notation. Recalling that

$\Theta^* = \begin{pmatrix} S_Y^* - L_Y^* & \Theta_{YX}^* \\ \Theta_{YX}^{*'} & \Theta_X^* \end{pmatrix}$  is the population precision matrix, let  $\tau_Y$  denote the minimum nonzero entry in magnitude of  $S_Y^*$ , let  $\deg(S_Y^*)$  denote the maximal number of nonzeros per row/column of  $S_Y^*$  (i.e., the degree of the graphical model underlying the conditional distribution  $Y|f(X), \zeta$ , which is specified by the precision matrix  $S_Y^*$ ), let  $\sigma_Y$  denote the minimum nonzero singular value of  $L_Y^*$ , and let  $\sigma_{YX}$  denote the minimum nonzero singular value of  $\Theta_{YX}^*$ .

**Theorem 2.1.** *Suppose we are given i.i.d observations  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n \subset \mathbb{R}^{p+q}$  of a collection of jointly Gaussian covariates/responses with population precision matrix  $\Theta^* \in \mathbb{S}_{++}^{p+q}$ . Fix  $\alpha > 0, \nu \in (0, 1/3), \omega_Y \in (0, 1), \omega_{YX} \in (0, 1)$ . Suppose the parameters  $\delta$  and  $\gamma$  are chosen such that the population Fisher information  $\mathbb{I}^*$  satisfies Assumptions 1 and 2.*

Let  $m \triangleq \max\{\frac{1}{\delta}, 1, \frac{1}{\gamma}\}$ ,  $\bar{m} \triangleq \max\{\delta, 1, \gamma\}$ ,  $\beta \triangleq \frac{3-\nu}{\nu}$ , and  $\psi \triangleq \|(\Theta^*)^{-1}\|_2$ . Further,  $C_1 = \frac{24}{\alpha} + \frac{1}{\psi^2}$ ,  $C_2 = \frac{4}{\alpha}(\frac{1}{3\beta} + 1)$ ,  $C_{\sigma_Y} = C_1^2\psi^2 \max\{12\beta + 1, \frac{2}{C_2\psi^2} + 1\}$ ,  $C_{\sigma_{YX}} = C_1^2\psi^2 \max\{18\beta, \frac{2}{C_2\psi^2} + 6\beta\}$ ,  $C_{s\text{amp}} = \max\{\frac{1}{48\psi\beta}, 48\beta\psi^3 C_1^2, 8\psi C_2, \frac{128\psi^3 C_2}{\alpha}\}$ , and  $\lambda_{\text{upper}} = \frac{1}{m\bar{m}^2 \deg(S_Y^*) C_{s\text{amp}}}$ . Suppose that the following conditions hold:

1.  $n \geq \frac{4608\psi^2\beta^2 m^2(p+q)}{\lambda_{\text{upper}}^2}$ ; that is  $n \gtrsim \left[\frac{\beta^4}{\alpha^2} m^4 \bar{m}^4 \deg(S_Y^*)^2\right](p+q)$
2.  $\lambda_n \in \left[\sqrt{\frac{4608\psi^2\beta^2 m^2(p+q)}{n}}, \lambda_{\text{upper}}\right]$ ; e.g.  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$
3.  $\tau_Y \geq 2C_1\delta\lambda_n$ ; that is  $\tau_Y \gtrsim \frac{\beta}{\alpha}\delta m \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$
4.  $\sigma_Y \geq \frac{m}{\omega_Y} C_{\sigma_Y} \lambda_n$ ; that is  $\sigma_Y \gtrsim \frac{\beta^2}{\alpha^2 \omega_Y} m \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$
5.  $\sigma_{YX} \geq \frac{m^2}{\omega_{YX}} C_{\sigma_{YX}} \gamma^2 \lambda_n$ ; that is  $\sigma_{YX} \gtrsim \frac{\beta^2 \gamma^2}{\alpha^2 \omega_{YX}} m^2 \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$

Then with probability greater than  $1 - 2\exp\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\}$ , the optimal solution  $(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y)$  of (1.6) with the observations  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n$  satisfies the following properties:

1.  $\text{sign}(\hat{S}_Y) = \text{sign}(S_Y^*)$ ,  $\text{rank}(\hat{L}_Y) = \text{rank}(L_Y^*)$ ,  $\text{rank}(\hat{\Theta}_{YX}) = \text{rank}(\Theta_{YX}^*)$
2.  $\Phi_{\delta,\gamma}(\hat{S}_Y - S_Y^*, \hat{L}_Y - L_Y^*, \hat{\Theta}_{YX} - \Theta_{YX}^*, \hat{\Theta}_X - \Theta_X^*) \leq C_1 \lambda_n$ ; that is  $\|\hat{S}_Y - S_Y^*\|_{\ell_\infty} \lesssim \frac{\beta}{\alpha} m \delta \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{L}_Y - L_Y^*\|_2 \lesssim \frac{\beta}{\alpha} m \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{\Theta}_{YX} - \Theta_{YX}^*\|_2 \lesssim \frac{\beta}{\alpha} \gamma m \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{\Theta}_X - \Theta_X^*\|_2 \lesssim \frac{\beta}{\alpha} m \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$ .

Notice that condition 1 of Theorem 2.1 ensures that the interval in condition 2 is non-empty. We give a proof of Theorem 2.1 in Appendix 5.1. Under the assumptions of the theorem, we construct appropriate primal feasible variables  $(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y)$  that satisfy the conclusions of the theorem – i.e.,  $\hat{\Theta}_{YX}, \hat{L}_Y$  are low-rank (with the same ranks as the underlying population quantities  $\Theta_{YX}^*$  and  $L_Y^*$ ) and  $\hat{S}_Y$  is sparse (with the same support as the underlying population quantity  $S_Y^*$ ) – and for which there exists a corresponding dual variable certifying optimality. This proof technique is sometimes also referred to as a primal-dual witness or certificate approach (Wainwright, 2009). The quantities  $\alpha, \beta$  (related to  $\nu$ , as stated in the theorem, via  $\beta = \frac{3-\nu}{\nu}$ ),  $\omega_Y, \omega_{YX}, \deg(S_Y^*)$  as well as the choices of the parameters  $\delta, \gamma$  play a prominent role in our result. Indeed, larger values of  $\alpha$  and  $\nu$  (leading to a better conditioned Fisher information, from Assumptions 1 and 2 in (2.5) and (2.6)) lead to less stringent requirements on the sample complexity, on the minimum magnitude nonzero entry  $\tau_Y$  of  $S_Y^*$ , on the minimum nonzero singular value  $\sigma_Y$  of  $L_Y^*$ , and on the minimum nonzero singular value  $\sigma_{YX}$  of  $\Theta_{YX}^*$ . In a similar vein, larger values of the quantities  $\omega_Y$  and  $\omega_{YX}$  in Assumptions 1 and 2 imply that the Fisher information is well-conditioned even for large distortions of the tangent spaces  $T(L_Y^*)$  and  $T(\Theta_{YX}^*)$  (see (2.3)), which in turn lead to less stringent requirements on the minimum nonzero singular values  $\sigma_Y$  and  $\sigma_{YX}$ .

As is clear from Theorem 2.1, the tradeoff parameters  $\delta, \gamma$  must be chosen such that the population Fisher information  $\mathbb{I}^*$  satisfies Assumptions 1 and 2 (see (2.5) and (2.6)) for some  $\alpha > 0, \nu \in (0, 1/3), \omega_Y > 0, \omega_{YX} > 0$ . Recall that these assumptions are stated in terms of the norm  $\Phi_{\delta,\gamma}$ , as defined in (2.4). The key complication that arises in characterizing values of  $\delta, \gamma$  for

which Assumptions 1 and 2 hold is that these parameters appear as multiplicative factors on norms imposed on different sub-blocks of matrices in  $\mathbb{S}^{p+q}$  (see the definition of  $\Phi_{\delta,\gamma}$  in (2.4)), which lead to conditions on gains of the Fisher information that are coupled across the different sub-blocks of  $\mathbb{S}^{p+q}$ . In order to overcome this difficulty, we describe in Appendix 5.2 a set of conditions on gains of the Fisher information restricted to the tangent spaces  $\Omega(S_Y^*), T_Y^*, T_{YX}^*$  *separately*; under these separable conditions, we explicitly characterize a non-empty polyhedral set  $V(\alpha, \nu, \omega_Y, \omega_{YX}) \subset \mathbb{R}^2$  such that Assumptions 1 and 2 hold for all  $(\delta, \gamma) \in V(\alpha, \nu, \omega_Y, \omega_{YX})$ . These conditions are interpretable and are stated in terms of the degree of the graphical model structure underlying the conditional distribution of  $S_Y^*$  (this quantity,  $\deg(S_Y^*)$ , makes an appearance in Theorem 2.1) and an incoherence parameter associated with the low-rank matrix  $L_Y^*$ .

### 3 Experiments

We illustrate the performance of the estimators corresponding to the SDR-FM, SDR-GM, and SDR-LVGM approaches (recall that these estimators are given by (1.2) with the choices (1.3), (1.4), and (1.5) for the regularizer  $R(\Theta_Y)$ ) and the estimator (1.7) in statistical modeling tasks involving financial asset data and newsgroup data. We solve these convex programs numerically using the LogdetPPA package developed for log-determinant semidefinite programs (Toh et al., 2002). Below we discuss the details of each dataset:

**financial:** We consider a financial asset modeling problem in which the responses  $Y$  are a collection of monthly stock returns of 67 companies from the Standard and Poor index from 1990 to 2005 and the covariates  $X$  are the following 7 variables –  $X_1$ : EUR to USD exchange rate,  $X_2$ : inflation rate,  $X_3$ : oil exports,  $X_4$ : industrial production index,  $X_5$ : population growth rate,  $X_6$ : consumer price index, and  $X_7$ : unemployment rate. Thus, we observe  $n = 188$  samples jointly of  $(Y, X) \in \mathbb{R}^{67} \times \mathbb{R}^7$ .

**newsgroup:** This dataset consists of  $n = 16242$  observations in  $\mathbb{R}^{100}$ , with each observation corresponding to a news document. The coordinates of these observations are indexed by a collection of 100 words, and each observation is a binary vector specifying whether a word appears in a document. Of these 100 words, the following 9 words are designated as covariates  $X \in \mathbb{R}^9$  as they are useful in categorizing newsgroup documents:  $X_1$ : government,  $X_2$ : religion,  $X_3$ : science,  $X_4$ : technology,  $X_5$ : war,  $X_6$ : medicine,  $X_7$ : world,  $X_8$ : food, and  $X_9$ : games. The remaining 91 words specify the response  $Y \in \mathbb{R}^{91}$ .

#### 3.1 Sufficient Dimension Reduction and Conditional Modeling

In this section, we investigate the performance of SDR-GM and SDR-LVGM (with the estimator (1.2) and choices (1.4) and (1.5) for the regularizer  $R(\Theta_Y)$ ) on the financial dataset. To illustrate the utility of incorporating information about the covariates, we also contrast these methods with two modeling approaches that do not account for the impact of the covariates  $X$  on the responses  $Y$  – we fit a sparse graphical model (denoted GM) as well as a latent-variable graphical model (denoted LVGM) to the responses  $Y$  using the Graphical Lasso technique (Yuan & Lin, 2006; Friedman et al., 2008) and the approach described by Chandrasekaran et al. (2012), respectively.

To properly compare the performance of these different techniques, we ensure that the complexity of each of the resulting models (in terms of the number of parameters required for their

specification) is approximately the same.<sup>3</sup> We begin by choosing a regularization parameter for the Graphical Lasso method such that small changes in the value of this parameter do not lead to substantial structural changes in the estimated graphical model, i.e., the estimated model is stable with respect to small changes in the regularization parameter. Following this approach, we use the GM approach on the financial dataset (without the covariates), and the resulting graphical model consists of 909 parameters (842 edges plus 67 node parameters). Next, we choose regularization parameters for the estimators corresponding to the LVGM (without covariates), SDR-GM (with both covariates and responses), and SDR-LVGM (with both covariates and responses) approaches such that the resulting models obtained via each of these techniques consist of approximately 909 parameters. Specifically, the model obtained using LVGM consists of 10 latent variables and a conditional graphical model (conditioned on the 11 latent variables) with 221 edges for a total of 913 parameters. The model obtained using SDR-GM consists of a 4-dimensional projection of the 7 covariates (that is sufficient with respect to the responses) and a conditional graphical model over the responses (conditioned on the dimension-reduced covariates) that consists of 564 edges for a total number of 908 parameters. Finally, the model obtained using SDR-LVGM consists of a 3-dimensional projection of the covariates, 7 latent variables, and a conditional graphical model (conditioned on the dimension-reduced covariates and the 7 latent variables) with 180 edges for a total of 908 parameters. Let  $\hat{\Theta}_{\text{GM}} \in \mathbb{S}_{++}^{67}$ ,  $\hat{\Theta}_{\text{LVGM}} \in \mathbb{S}_{++}^{67}$ ,  $\hat{\Theta}_{\text{SDR-GM}} \in \mathbb{S}_{++}^{67+7}$ , and  $\hat{\Theta}_{\text{SDR-LVGM}} \in \mathbb{S}_{++}^{67+7}$  denote the precision matrices of the models corresponding to GM, LVGM, SDR-GM, and SDR-LVGM respectively. Although these models have similar complexities, they have different predictive performances, as described next.

We assess the predictive performance of each of these four models on 90 monthly observations  $\{(Y_{\text{test}}^{(j)}, X_{\text{test}}^{(j)})_{j=1}^{90} \in \mathbb{R}^{67} \times \mathbb{R}^7$  from 2006 to 2013 (recall that the training set based on which the four models were obtained consisted of 188 monthly observations during the period 1990 to 2006). For the models obtained via GM and LVGM, we compute the average log-likelihood over the test samples using the distributions specified by the precision matrices  $\hat{\Theta}_{\text{GM}}$  and  $\hat{\Theta}_{\text{LVGM}}$  respectively. For the model obtained via SDR-GM that accounts for the influence of the covariates on the responses, we compute the log-likelihood of each test sample  $Y_{\text{test}}^{(j)}$  (and subsequently average over all test samples) with respect to the distribution of  $Y|[f(X) = f(X_{\text{test}}^{(j)})]$  where  $f(X)$  is the projection of  $X$  into the row-space of the matrix  $-(\hat{\Theta}_{\text{SDR-GM}})_{\bar{Y}}^{-1}(\hat{\Theta}_{\text{SDR-GM}})_{YX}$  (recall that  $-(\hat{\Theta}_{\text{SDR-GM}})_{\bar{Y}}^{-1}(\hat{\Theta}_{\text{SDR-GM}})_{YX}$  denotes the map of best linear estimator of  $Y$  based on  $X$ ). We follow a similar approach to compute the predictive performance of the model obtained via SDR-LVGM. The average predictive log-likelihoods of models obtained via GM, LVGM, SDR-GM, and SDR-LVGM are  $-127.55$ ,  $-122.73$ ,  $-121.12$ , and  $-120.28$  respectively. For comparison, a model obtained via FM (factor modeling) on the training set – without incorporating the covariates – consists of 14 latent factors (for a total of 914 parameters) and gives an average predictive log-likelihood of  $-123.99$  over the test set. On the other hand, a model obtained via SDR-FM (using the estimator (1.2) with the regularizer  $R(\Theta_Y)$  set as in (1.3)) on the training set – that incorporates observations of both the covariates and the responses – consists of a 5-dimensional projection of the covariates and 8 latent factors in the conditional model of the responses given the dimension-reduced covariates (the total number of parameters equals 920), and it provides an average predictive log-likelihood of  $-121.35$  on the test set. As larger values of average log-likelihood are indicative of a better fit to the test samples, these results suggest that SDR-LVGM offers the best predictive performance of the different approaches

<sup>3</sup>The number of parameters required to specify a sparse graphical model with a precision matrix  $N \in \mathbb{S}^p$  is equal to  $p$  plus one-half the number of nonzero off-diagonal entries of  $N$  (as  $N$  is symmetric). The number of parameters required to specify a  $p \times q$  matrix with rank  $r \leq \min\{p, q\}$  is equal to  $r(p+q) - r^2$ . Finally, the number of parameters required to specify a matrix in  $\mathbb{S}^p$  with rank  $r \leq p$  is equal to  $rp - r(r-1)/2$ .

considered in this experiment.

Focussing on the structural aspects of the models obtained via GM, LVGM, SDR-GM, and SDR-LVGM, Fig. 1 shows the (conditional) graphical model structures over the responses corresponding to each of these approaches. The ten strongest edges in the conditional graphical model obtained via SDR-LVGM in Fig. 1(d) (in terms of the magnitude of the entries in the precision matrix) are between General Electric - American Express, Target - Bancorp, Hewlett Packard - Oracle, Texas Instruments - General Electric, Occidental Petroleum Corp. - Wells Fargo, American Insurance Group - Bank of New York, Merck & Co. - Walgreens, JP Morgan - Verizon, Verizon - CVS Health, Pfizer Inc. - Colgate. The presence of some strong edges between entities in different industries suggests that dependencies between assets belonging to the same industry may be better modeled via the latent variables or the dimension-reduced covariates. Each of these ten edges in the conditional graphical model obtained via SDR-LVGM also appears as a strong edge in the conditional graphical model obtained via LVGM (in which the covariates are not incorporated). Examples of other strong edges in the conditional graphical model obtained via LVGM (which do not appear in the conditional graphical model obtained via SDR-LVGM) include IBM Corp. - American Insurance Group, Hewlett Packard - Southern Company, Hewlett Packard - Walmart, General Electric - Colgate, Fedex - Emerson.

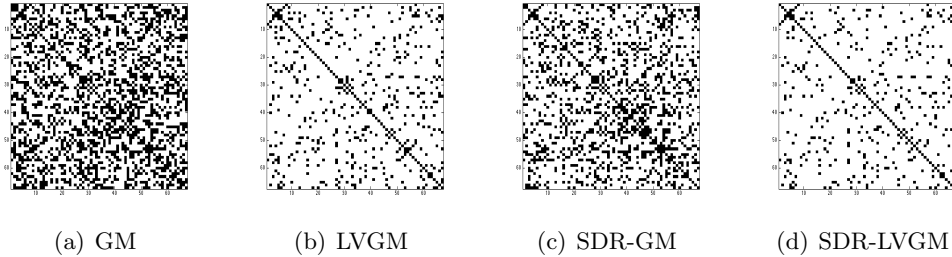


Figure 1: These figures show the sparsity pattern (black denotes an edge, and white denotes no edge) of the graphical models associated with each modeling paradigm

Turning our attention to the latent components identified in the models obtained via LVGM and SDR-LVGM, we note that the number of latent variables in the model obtained using SDR-LVGM (7 variables) is smaller than that in the model obtained via LVGM (10 variables). This observation suggests that the 3-dimensional projection of the covariates in the model obtained via SDR-LVGM accounts for some of the effect of the latent variables in the model obtained using LVGM (which does not incorporate the covariates). To illuminate this point quantitatively, let the matrix  $\Gamma = -(\hat{\Theta}_{\text{SDR-LVGM}})_Y^{-1}(\hat{\Theta}_{\text{SDR-LVGM}})_{YX} \in \mathbb{R}^{67 \times 7}$  denote the map of the best linear estimator of  $Y$  based on  $X$  (specified by the model obtained via SDR-LVGM) and let the matrix  $\Lambda \in \mathbb{R}^{67 \times 10}$  denote the map of best linear estimator of  $Y$  based on the latent variables  $\zeta \in \mathbb{R}^{10}$  (specified by the model obtained via LVGM<sup>4</sup>). The column space of  $\Gamma$  corresponds to the 3-dimensional image of the mapping of the best linear estimator of  $Y$  based on  $X$ , and it represents the component of the response  $Y$  that is correlated with the covariates  $X$  in the model obtained using the SDR-LVGM approach. Similarly, the column space of  $\Lambda$  represents the 10-dimensional component of  $Y$  that is correlated with the latent variables  $\zeta$  in the model obtained using the LVGM approach. As such, the closeness between these column spaces measures the degree to which the sufficient dimension reduction  $f(X)$  accounts for some of the influence of the latent variables  $\zeta$  on the covariates  $Y$ . The largest principal angle between the 3-dimensional column space of  $\Gamma$  and the 10-dimensional column space of  $\Lambda$  is 11.06 degrees. This result indicates that the dimension reduced covariates in

<sup>4</sup>The matrix  $\Lambda$  is only known up to right-multiplication by a non-singular linear transformation. As a result, the key invariants are the rank and the column space of  $\Lambda$ .

the model obtained using SDR-LVGM account for some of the effect of the latent components in the model obtained via LVGM.

### 3.2 Combining Covariate Selection and Graphical Modeling

In this section, we investigate the performance of the estimator (1.7) on the financial and the newsgroup datasets. Recall that this estimator selects a subset of the covariates  $X$  that is most useful for predicting the responses  $Y$ , while simultaneously fitting a latent-variable graphical model to the conditional distribution of the responses given selected covariates. We denote this modeling approach as CS-LVGM.

We apply the estimator (1.7) to the financial and the newsgroup datasets with fixed choices for parameters  $\lambda_n, \delta$  and different choices of the parameter  $\gamma$ . The objective of this experiment is to illustrate the different covariates selected in each dataset as  $\gamma$  is varied. For each of the models obtained using this CS-LVGM approach, Table 1 and Table 2 list the subset of covariates that were selected (recall that the selected covariates are represented by the indices of the nonzero columns of the submatrix  $\hat{\Theta}_{YX} \in \mathbb{R}^{p \times q}$  of the estimated joint precision matrix  $\hat{\Theta}_{++}^{p+q}$ ). As expected, larger values of  $\gamma$  yield a smaller subset of relevant covariates as the regularization term  $\gamma \|\Theta_{YX}\|_{1,2}$  in (1.7) is enforced more strongly. With the financial dataset, the covariates  $X_1$  (population growth rate) and  $X_5$  (EUR to USD exchange rate) persist as  $\gamma$  increases, which suggests that they are the most relevant for predicting stock returns among the seven covariates considered in the experiment. On the other hand, the covariate  $X_3$  (oil exports) appears to be the least influential. For the newsgroup dataset, the covariates  $X_1$  (government) and  $X_7$  (world) persist as  $\gamma$  increases, suggesting that these are the most useful for predicting word occurrences in documents in the 20newsgroup dataset, while the covariates  $X_6$  (medicine) and  $X_8$  (food) do not seem as relevant.

$\gamma$	covariates identified
1.50	$\{X_1, X_2, X_4, X_5, X_6, X_7\}$
1.99	$\{X_1, X_2, X_4, X_5\}$
2.14	$\{X_1, X_5, X_7\}$
2.57	$\{X_1, X_5\}$

Table 1:  $\gamma$  vs selected covariates for financial dataset with  $\lambda_n = 0.58$  and  $\delta = 0.29$

$\gamma$	covariates identified
1.39	$\{X_1, X_2, X_3, X_4, X_5, X_7, X_9\}$
2.32	$\{X_1, X_2, X_3, X_5, X_9\}$
2.80	$\{X_1, X_2, X_3, X_9\}$
3.02	$\{X_1, X_7\}$

Table 2:  $\gamma$  vs selected covariates for newsgroup dataset with  $\lambda_n = 0.38$  and  $\delta = 0.41$

We inspect more closely the model obtained using the CS-LVGM approach on the newsgroup dataset with parameters  $\lambda_n = 0.38, \delta = 0.41, \gamma = 2.32$  (this corresponds to the second line in Table 2). This model consists of 5 selected covariates, 6 latent variables, and a conditional graphical model (conditioned on the 5 covariates and the 6 latent variables) with 10 edges for a total number of 1087 parameters. The 10 edges in this conditional graphical model are between God-Jesus, Dos-Windows, Bible-God, card-video, email-phone, Christian-God, state-university, computer-university, disk-drive, and Israel-Jews. For comparison, we obtain a model using the LVGM approach (ignoring the covariates) of comparable complexity with a total of 1083 parameters – this model consists of 11 latent variables and a conditional graphical model (conditioned on the 11 latent variables) with 46 edges. Each of the ten edges in the conditional graphical model obtained via CS-LVGM appear as stronger edges in the conditional graphical model obtained via LVGM. Examples of additional strong edges in the conditional graphical model obtained via LVGM include patients-disease, hockey-NHL, and players-baseball. The edge between patients-disease presents an interesting illustration as these two words appear together in 64 documents. In 45 of these 64 documents, however, at least one of the 5 covariates selected by the CS-LVGM model ( $X_1, X_2, X_3, X_5, X_9$  from Table 2) also appears. Thus, the lack of an edge between patients-disease

in the conditional graphical model obtained using the CS-LVGM approach is perhaps explained by the inclusion of the the 5 covariates. In a similar vein, the absence of the edges hockey-NHL and players-baseball in the conditional graphical model obtained via CS-LVGM may be attributed to the inclusion of the covariate  $X_9$  (game). In contrast, a curious point arises when considering the presence of the edges God-Jesus, Bible-God, Christian-God, and Israel-Jews in the conditional graphical model obtained via CS-LVGM. From a preliminary inspection, one might expect that the inclusion of the covariates  $X_2, X_7$  (religion, world) would account for these four pairs of interactions, resulting in their absence in the conditional graphical model. However, (as an example) the words God and Jesus appear together in 380 documents, and only 180 of these documents include at least one of the 5 selected covariates. Thus, it is not surprising that the edge God-Jesus remains in the conditional graphical model obtained using the CS-LVGM approach despite this graphical model being conditioned on the 5 selected covariates.

## 4 Further Directions

The selection of regularization parameters is a common practical challenge in high-dimensional estimation problems. Approaches based on cross-validation are widely used, although these techniques optimize for prediction performance and do not always yield concise models. To address this shortcoming, several methods have been proposed recently for the selection of regularization parameters for the Lasso (Meinshausen & Bühlmann, 2010). It is of interest to extend these ideas to our estimator (1.2), which requires the specification of multiple regularization parameters. Further, the convex programs proposed in this paper are computationally tractable (i.e., solvable to a desired accuracy in polynomial time), but it would be useful to develop special-purpose numerical schemes – perhaps building on recent scalable algorithms for the Graphical Lasso (Friedman et al., 2008; Hsieh et al., 2013) – for efficient solution in massive-scale problems. Finally, it is of interest to extend our techniques to non-Gaussian settings, in which one may be interested in identifying nonlinear dimension reductions of the covariates that are sufficient with respect to the responses.

## References

- ADRAGNI, K. P & COOK, D. R. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of Royal Society A*. **367**, 4385–4400.
- BACH, F. (2008). Consistency of trace norm minimization. *Journal of Machine Learning Research*. **9**, 1019–1048.
- BICKEL, P. J. & LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*. **36**, 199–227.
- BICKEL, P. J. & LEVINA, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*. **36**, 2577–2604.
- BREM, R. & KRUGLYAK, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceeding of National Academy of Sciences*. **102**, 1572–1577.
- BÜHLMANN, P & VAN DE GEER, S. (2011). *Statistics for high-dimensional data*. New York: Springer.



- CANDÈS, E. J., ROMBERG, J. & TAO, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transaction on Information Theory*. **52**, 489–509.
- CANDÈS, E. J. & RECHT, B. (2009). Exact matrix completion via convex optimization. *Foundation of Computational Mathematics*. **9**, 717–772.
- CAI, T., LI, H., LIU, W. & XIE, J. (2010). Covariate adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*. **106**, 1–19.
- CHANDRASEKARAN, V., PARRILO, P. A. & WILLSKY, A. S. (2012). Latent Variable Graphical Model Selection via Convex Optimization. *Annals of Statistics*. **40**, 1935–1967.
- CHECHIK, G., GLOBERSON, A., TISHBY, N. & WEISS, Y. (2005). Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*. **6**, 165–188.
- CHEN, S. S., DONOHO, D. L. & SAUNDERS, M. A (2001). Atomic decomposition by basis pursuit. *SIAM Review*. **43**, 129–159.
- CHEUNG, V. & SPIELDMAN, R. (2002). The genetics of variation in gene expression. *Nature Genetics*. **35**, 2313–2351.
- COOK, R. D. & NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of American Statistical Association*. **100**, 410–428.
- COOK, R. D. & WEISBERG, S. (1991). Discussion of “Sliced inverse regression for dimension reduction”. *Journal of American Statistical Association*. **86**, 328–332.
- DAVIDSON, K.R. & SZAREK, S.J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces*. **I**, 317–366.
- DONOHO, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*. **52**, 1289–1306.
- EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*. **36** 2717–2756.
- DUAN, N. & LI, K.C. (1991). Slicing regression: A link-free regression method. *Annals of Statistics*. **19**, 505–530.
- FAN, J., FAN, Y. & LV, J. (2008). High-dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*. **147** 186–197.
- FAZEL, M. (2002). Matrix rank minimization with applications. *PhD thesis, Department of Electrical Engineering, Stanford University*. 2002.
- FRIEDMAN, J., HASTIE, T & TIBSHIRANI, R (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. **9** 432–441.
- FUNG, W. K., HE, X., LIU, L. & SHI, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*. **12** 1093–1113.
- HORN, R. A. & JOHNSON, C. R. (1990). Matrix Analysis. *Cambridge University Press, Cambridge*.

- HSIEH, C.-J., SUSTIK, M. A., DHILLON, I. S., & RAVIKUMAR, P. (2013). BIG & QUIC: Sparse inverse covariance estimation for million variables. *Neural Information Processing System*.
- KATO, T. (1995). Perturbation theory for linear operators. *Springer*.
- LI, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of American Statistical Association*. **86** 316–342.
- LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein’s lemma. *Journal of American Statistical Association*. **87** 1025–1039.
- MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*. **34** 1436–1462.
- MEINSHAUSEN, N. & BÜHLMANN, P. (2010). Stability selection. *Journal of Royal Statistical Society: Series B*. **34** 1436–1462.
- NEGAHBAN, S. & WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*. **39** 1069–1097.
- RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. & YU, B. (2011). High-dimensional covariance estimation by minimizing  $\ell_1$ -penalized log-determinant divergence. *Electronic Journal of Statistics* **4** 935–980.
- RECHT, B., FAZEL, M. & PARRILO, P. A. (2010). Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review* **52** 471–501.
- ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*. **2** 494–515.
- SOHN, K.-A. & KIM, S. (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. *Proceeding 15th International Conference Artificial Intelligence and Statistics*. **14** 1081–1089.
- TIBSHIRANI, R (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistics Society*. **58** 267–288.
- TOH, K. C, TODD, M. J. & TUTUNCU, R. H. *SDPT3 - a MATLAB software package for semidefinite-quadratic-linear programming*. Available from <http://www.math.nus.edu.sg/~mattohc/sdpt3.html>.
- WAINWRIGHT, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using  $\ell_1$ -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*. **55** 2183–2202.
- WAINWRIGHT, M. J. (2014). Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and its Applications*. **1** 233–253.
- YIN, J. & LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of general genomic data. *Annals of Applied Statistics*. **5** 2630–2650.
- YUAN, M.(2008) Efficient computation of  $\ell_1$  regularized estimates in Gaussian graphical models. *Journal of Computation and Graphical Statistics*. **17** 809–826.

- YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society: Series B.* **68** 49–67.
- YUAN, M. & LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika.* **94** 19–35.
- YUAN, X. & ZHANG, T. (2014). Partial Gaussian graphical model estimation. *IEEE Transactions on Information Theory.* **60** 1673–1687.
- ZHAO, P. & YU, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research.* **7** 2541–2567.

## 5 Appendix

### 5.1 Proof of Theorem 2.1

In this section, we prove the consistency results (stated in Theorem 2.1) of the estimator (1.6). The high-level proof strategy is similar in spirit to the proof of the consistency results for sparse graphical model recovery (Ravikumar et al., 2008) and latent variable graphical model recovery (Chandrasekaran et al., 2012). However, the estimator (1.6) is different than the estimators proposed by (Ravikumar et al., 2008) and (Chandrasekaran et al., 2012) due to the nuclear norm penalty corresponding to the SDR objective.

We begin by considering the following convex optimization program:

$$\begin{aligned}
 (\bar{\Theta}, \bar{S}_Y, \bar{L}_Y) = & \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \Theta \succ 0 \\ S_Y, L_Y \in \mathbb{S}^p}}{\operatorname{argmin}} & -\ell(\Theta; \{X^{(i)}, Y^{(i)}\}_{i=1}^n) + \lambda_n[\delta \|S_Y\|_{\ell_1} + \|L_Y\|_{\star} + \gamma \|\Theta_{YX}\|_{\star}] \\
 \text{s.t.} & & \Theta_Y = S_Y - L_Y
 \end{aligned} \tag{5.1}$$

Comparing (5.1) with the convex program (1.6), the difference is that we no longer constrain  $L_Y$  to be a positive semidefinite matrix. In particular, if  $L_Y \succeq 0$ , then the nuclear norm of the matrix  $L_Y$  in the objective function of (5.1) reduces to the trace of  $L_Y$ . We show that the unique optimum  $(\bar{\Theta}, \bar{S}_Y, \bar{L}_Y)$  of (5.1) has the property that with high probability,  $\bar{L}_Y$  is positive semidefinite. As a result, with high probability, the variables  $(\bar{\Theta}, \bar{S}_Y, \bar{L}_Y)$  are also the optimum of (1.6). In the remainder of this section, we show that under the assumptions of Theorem 2.1, the primal feasible variables  $(\bar{\Theta}, \bar{S}_Y, \bar{L}_Y)$  are structurally correct estimates of  $(\Theta^*, S_Y^*, L_Y^*)$  (see Definition 2.1). Below, we outline our proof strategy:

1. We proceed by analyzing (5.1) with additional constraints that the variables  $S_Y$ ,  $L_Y$ , and  $\Theta_{YX}$  belong to the algebraic varieties of sparse and low-rank matrices (specified by the support of  $S_Y^*$  and rank of  $L_Y^*$  and  $\Theta_{YX}^*$ ), and that the tangent spaces  $\Omega(S_Y)$ ,  $T(L_Y)$ ,  $T(\Theta_{YX})$  are close to the nominal tangent spaces  $\Omega(S_Y^*)$ ,  $T(L_Y^*)$ , and  $T(\Theta_{YX}^*)$  respectively. We prove that under suitable conditions on the minimum magnitude nonzero entry of  $S_Y^*$ , minimum nonzero singular value of  $L_Y^*$ , and minimum nonzero singular value of  $\Theta_{YX}^*$ , any optimum set of variables  $(\Theta, S_Y, L_Y)$  of this non-convex program are smooth points of the underlying varieties; that is  $\operatorname{sign}(S_Y) = \operatorname{sign}(S_Y^*)$ ,  $\operatorname{rank}(L_Y) = \operatorname{rank}(L_Y^*)$  and  $\operatorname{rank}(\Theta_{YX}) = \operatorname{rank}(\Theta_{YX}^*)$ . Further, we show that  $L_Y$  has the same inertia as  $L_Y^*$  so that  $L_Y \succeq 0$ .
2. Conclusions of the previous step imply the the variety constraints can be “linearized” at the global optimum of the non-convex program to obtain tangent-space constraints. Under suitable conditions on the regularization parameter  $\lambda_n$ , we prove that with high probability, the unique optimum of this “linearized” program coincides with the global optimum of the non-convex program.
3. Finally, we show that the tangent-space constraints of the linearized program are inactive at the optimum. Therefore, the optimal solution of (5.1) has the property that with high probability:  $\operatorname{sign}(\bar{S}_Y) = \operatorname{sign}(S_Y^*)$ ,  $\operatorname{rank}(\bar{L}_Y) = \operatorname{rank}(L_Y^*)$ , and  $\operatorname{rank}(\bar{\Theta}_{YX}) = \operatorname{rank}(\Theta_{YX}^*)$ . Since  $\bar{L}_Y \succeq 0$ , we conclude that the variables  $(\bar{\Theta}, \bar{S}_Y, \bar{L}_Y)$  are the unique optimum of (1.6).

In Section 5.1.1, we prove the results of step 1. In Section 5.1.2, we prove the results of step 2. Finally, in Section 5.1.3, we prove the results of step 3.

### 5.1.1 Variety Constrained Optimization Program

Letting  $m \triangleq \max\{\frac{1}{\delta}, 1, \frac{1}{\gamma}\}$  and  $\psi \triangleq \|(\Theta^*)^{-1}\|_2$ , we consider the following variety-constrained optimization program:

$$\begin{aligned} (\Theta^{\mathcal{M}}, S_Y^{\mathcal{M}}, L_Y^{\mathcal{M}}) = & \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \Theta \succ 0 \\ S_Y, L_Y \in \mathbb{S}^p}}{\operatorname{argmin}} & -\ell(\Theta; \{X^{(i)}, Y^{(i)}\}_{i=1}^n) + \lambda_n[\delta\|S_Y\|_{\ell_1} + \|L_Y\|_* + \gamma\|\Theta_{YX}\|_*] \\ \text{s.t.} & \Theta_Y = S_Y - L_Y, (\Theta, S_Y, L_Y) \in \mathcal{M}. \end{aligned} \quad (5.2)$$

Here, the set  $\mathcal{M}$  is given by:

$$\begin{aligned} \mathcal{M} \triangleq & \left\{ (\Theta, S_Y, L_Y) \in \mathbb{S}^{(p+q)} \times \mathbb{S}^p \times \mathbb{S}^p \mid S_Y \in \Omega(S_Y^*), \operatorname{rank}(L_Y) \leq \operatorname{rank}(L_Y^*) \right. \\ & \operatorname{rank}(\Theta_{YX}) \leq \operatorname{rank}(\Theta_{YX}^*); \\ & \|\mathcal{P}_{T(L_Y^*)^\perp}(L_Y - L_Y^*)\|_2 \leq \frac{\omega_Y \lambda_n}{2m\psi^2}, \\ & \|\mathcal{P}_{T(\Theta_{YX}^*)^\perp}(\Theta_{YX} - \Theta_{YX}^*)\|_2 \leq \frac{\omega_{YX} \lambda_n}{2m\psi^2} \\ & \left. \Phi_{\delta, \gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \Delta] \leq 5\lambda_n \right\} \end{aligned}$$

The optimization program (5.2) is non-convex due to the rank constraints  $\operatorname{rank}(L_Y) \leq \operatorname{rank}(L_Y^*)$  and  $\operatorname{rank}(\Theta_{YX}) \leq \operatorname{rank}(\Theta_{YX}^*)$  in the set  $\mathcal{M}$ . These constraints, in addition to the constraint  $S_Y \in \Omega(S_Y^*)$  ensure that the matrices  $S_Y, L_Y$ , and  $\Theta_{YX}$  belong to appropriate varieties. The constraints in  $\mathcal{M}$  along  $T(L_Y^*)^\perp$  and  $T(\Theta_{YX}^*)^\perp$  ensure that the tangent spaces  $T(L_Y)$  and  $T(\Theta_{YX})$  are “close” to  $T(L_Y^*)$  and  $T(\Theta_{YX}^*)$  respectively. Finally, the last condition roughly controls the error. In this section, we will prove that any feasible set of variables  $(\Theta, S_Y, L_Y)$  – and in particular an optimal set of variables  $(\Theta^{\mathcal{M}}, S_Y^{\mathcal{M}}, L_Y^{\mathcal{M}})$  – is structurally correct estimate of  $(\Theta^*, S_Y^*, L_Y^*)$ . We begin by proving that any feasible set of variables  $(\Theta, S_Y, L_Y)$  is “close” in norm to the population quantities  $(\Theta^*, S_Y^*, L_Y^*)$ .

**Proposition 5.1.** *Let  $(\Theta, S_Y, L_Y)$  be a set of feasible variables of (5.2). Let  $\Delta = (S_Y - S_Y^*, L_Y - L_Y^*, \Theta_{YX} - \Theta_{YX}^*, \Theta_X - \Theta_X^*)$  and  $C_1 = \frac{12}{\alpha} + \frac{1}{\psi^2}$ . Then,  $\Phi_{\delta, \gamma}[\Delta] \leq C_1 \lambda_n$*

*Proof.* Let  $\mathbb{H}^* = \Omega(S_Y^*) \times T(L_Y^*) \times T(\Theta_{YX}^*) \times \mathbb{S}^q$ . Then,

$$\begin{aligned} \Phi_{\delta, \gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}^*}(\Delta)] & \leq \Phi_{\delta, \gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(\Delta)] + \Phi_{\delta, \gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}^{*\perp}}(\Delta)] \\ & \leq 5\lambda_n + m\psi^2 \left( \frac{\omega_Y \lambda_n}{2m\psi^2} + \frac{\omega_{YX} \lambda_n}{2m\psi^2} \right) \leq 6\lambda_n \end{aligned}$$

Since  $\Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}^*}(\cdot)] \leq 2\Phi_{\delta, \gamma}(\cdot)$ , we have that  $\Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}^*} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}^*}(\Delta)] \leq 12\lambda_n$ . Consequently, we appeal to the Fisher information Assumption 1 in (2.5) to conclude that  $\Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}^*}(\Delta)] \leq \frac{12\lambda_n}{\alpha}$ . Moreover:

$$\Phi_{\delta, \gamma}[\Delta] \leq \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}^*}(\Delta)] + \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}^{*\perp}}(\Delta)] \leq \frac{12\lambda_n}{\alpha} + \frac{\lambda_n}{\psi^2} = C_1 \lambda_n$$

□

Proposition 5.1 leads to powerful implications. In particular, under additional conditions on the minimum magnitude nonzero entry of  $S_Y^*$ , and minimum nonzero singular values of  $L_Y^*$  and  $\Theta_{YX}^*$ , any feasible set of variables  $(\Theta, S_Y, L_Y)$  of (5.2) has two key properties: (a) The variables

$(\Theta_{YX}, S_Y, L_Y)$  are smooth points of the underlying varieties, (b) The constraints in  $\mathcal{M}$  along  $T(L_Y^\star)^\perp$  and  $T(\Theta_{YX}^\star)^\perp$  are locally inactive at  $\Theta_{YX}$  and  $L_Y$ . These properties, among others, are proved in the following corollary.

**Corollary 5.2.** *Consider any feasible variables  $(\Theta, S_Y, L_Y)$  of (5.2). Let  $\sigma_Y$  be the smallest nonzero singular value of  $L_Y^\star$ ,  $\sigma_{YX}$  be the smallest nonzero singular value of  $\Theta_{YX}^\star$ , and  $\tau_Y$  the minimum magnitude nonzero element of  $S_Y^\star$ . Let  $\mathbb{H}' = \Omega(S_Y) \times T(L_Y) \times T(\Theta_{YX}) \times \mathbb{S}^q$  and  $C_{T'} = \mathcal{P}_{\mathbb{H}'^\perp}(0, L_Y^\star, \Theta_{YX}^\star, 0)$ . Furthermore, let  $C_1 = \frac{12}{\alpha} + \frac{1}{\psi^2}$ ,  $C_2 = \frac{4}{\alpha}(1 + \frac{1}{3\beta})$ ,  $C_{\sigma_Y} = C_1^2 \psi^2 \max\{12\beta + 1, \frac{2}{C_2 \psi^2} + 1\}$  and  $C'_{\sigma_{YX}} = C_1^2 \psi^2 \max\{12\beta + \frac{6\beta}{\gamma}, \frac{2}{C_2 \psi^2} + \frac{6\beta}{\gamma}\}$ . Suppose that the following inequalities are met:  $\sigma_Y \geq \frac{m}{\omega_Y} C_{\sigma_Y} \lambda_n$ ,  $\sigma_{YX} \geq \frac{m\gamma^2}{\omega_{YX}} C'_{\sigma_{YX}} \lambda_n$ , and  $\tau_Y \geq 2\delta C_1 \lambda_n$ . Then,*

1.  $L_Y$  and  $\Theta_{YX}$  are smooth points of their underlying varieties, i.e.  $\text{rank}(L_Y) = \text{rank}(L_Y^\star)$ ,  $\text{rank}(\Theta_{YX}) = \text{rank}(\Theta_{YX}^\star)$ ; Moreover  $L_Y$  has the same inertia as  $L_Y^\star$ .
2.  $\|\mathcal{P}_{T(L_Y^\star)^\perp}(L_Y - L_Y^\star)\|_2 \leq \frac{\lambda_n \omega_Y}{48m\psi^2}$  and  $\|\mathcal{P}_{T(\Theta_{YX}^\star)^\perp}(\Theta_{YX} - \Theta_{YX}^\star)\|_2 \leq \frac{\lambda_n \omega_{YX}}{48m\psi^2}$
3.  $\rho(T(L_Y), T(L_Y^\star)) \leq \omega_Y$ , and  $\rho(T(\Theta_{YX}), T(\Theta_{YX}^\star)) \leq \omega_{YX}$ ; that is, the tangent spaces at  $L_Y$  and  $\Theta_{YX}$  are “close” to the tangent spaces at  $L_Y^\star$  and  $\Theta_{YX}^\star$  respectively.
4.  $\Phi_{\delta, \gamma}[\mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} C_{T'}] \leq \frac{\lambda_n}{6\beta}$
5.  $\Phi_{\delta, \gamma}[C_{T'}] \leq C_2 \lambda_n$
6.  $S$  is the smooth point of its underlying variety, i.e.  $\text{sign}(S_Y) = \text{sign}(S_Y^\star)$

*Proof.* We note the following relations before proving each step:  $C_1 \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2}$ ,  $\omega_Y, \omega_{YX} \in [0, 1]$ , and  $\beta \triangleq \frac{3-\nu}{\nu} \geq 8$  for  $\nu \in (0, 1/3]$ . We also appeal to the results of (Kato, 1995; Bach, 2008; Chandrasekaran et al., 2012) regarding perturbation analysis of the low-rank matrix variety.

1. Based on the assumptions regarding the minimum nonzero singular values of  $L_Y^\star$  and  $\Theta_{YX}^\star$ , we have:

$$\begin{aligned} \sigma_Y &\geq \frac{C_1^2 \lambda_n}{\omega_Y} m \psi^2 (12\beta + 1) \geq \frac{C_1 \lambda_n}{\omega_Y} (12\beta + 1) \geq (12\beta + 1) C_1 \lambda_n \geq 8 C_1 \lambda_n \geq 8 \|L - L_Y^\star\|_2 \\ \sigma_{YX} &\geq \frac{C_1^2 \lambda_n}{\omega_{YX}} \gamma^2 m \psi^2 \left( \frac{6\beta}{\gamma} + 12\beta \right) \geq C_1^2 \lambda_n \gamma^2 m \psi^2 \frac{6\beta}{\gamma} \geq 8 \gamma C_1 \lambda_n \geq 8 \|\Theta_{YX} - \Theta_{YX}^\star\|_2 \end{aligned}$$

Combing these results and Proposition 5.1, we conclude that  $L_Y$  and  $\Theta_{YX}$  are smooth points of their respective varieties, i.e.  $\text{rank}(L_Y) = \text{rank}(L_Y^\star)$ , and  $\text{rank}(\Theta_{YX}) = \text{rank}(\Theta_{YX}^\star)$ . Furthermore,  $L_Y$  has the same inertia as  $L_Y^\star$ .

2. Since  $\sigma_Y \geq 8 \|L_Y - L_Y^\star\|_2$ , and  $\sigma_{YX} \geq 8 \|\Theta_{YX} - \Theta_{YX}^\star\|_2$ , we can appeal to Proposition 2.2 in (Chandrasekaran et al., 2012) to conclude that the constraints in  $\mathcal{M}$  along  $\mathcal{P}_{T(L_Y^\star)^\perp}$  and  $\mathcal{P}_{T(\Theta_{YX}^\star)^\perp}$  are strictly feasible:

$$\begin{aligned} \|\mathcal{P}_{T(L_Y^\star)^\perp}(L_Y - L_Y^\star)\|_2 &\leq \frac{\|L_Y - L_Y^\star\|_2^2}{\sigma_Y} = \frac{C_1^2 \lambda_n^2 \omega_Y}{C_1^2 \lambda_n m \psi^2 (12\beta + 1)} \leq \frac{\lambda_n}{48m\psi^2} \\ \|\mathcal{P}_{T(\Theta_{YX}^\star)^\perp}(\Theta_{YX} - \Theta_{YX}^\star)\|_2 &\leq \frac{\|\Theta_{YX} - \Theta_{YX}^\star\|_2^2}{\sigma_{YX}} = \frac{C_1^2 \gamma^2 \lambda_n^2 \omega_{YX}}{C_1^2 \lambda_n m \psi^2 \gamma^2 \left( \frac{6\beta}{\gamma} + 12\beta \right)} \leq \frac{\lambda_n}{48m\psi^2} \end{aligned}$$

3. Appealing to Proposition 2.1 in (Chandrasekaran et al., 2012), we prove that the tangent spaces  $T(L_Y)$  and  $T(\Theta_{YX})$  are close to  $T(L_Y^*)$  and  $T(\Theta_{YX}^*)$  respectively:

$$\begin{aligned}\rho(T(L_Y), T(L_Y^*)) &\leq \frac{2\|L_Y - L_Y^*\|_2}{\sigma_Y} \leq \frac{2C_1\lambda_n\omega_Y}{C_1^2\lambda_n m\psi^2(12\beta+1)} \leq \omega_Y \\ \rho(T(\Theta_{YX}), T(\Theta_{YX}^*)) &\leq \frac{2\|\Theta_{YX} - \Theta_{YX}^*\|_2}{\sigma_{YX}} \leq \frac{2\lambda_n C_1\gamma\omega_{YX}}{C_1^2\lambda_n m\psi^2\gamma^2\frac{6\beta}{\gamma}} \leq \omega_{YX}\end{aligned}$$

4. Letting  $\sigma'_Y$  and  $\sigma'_{YX}$  be the minimum nonzero singular value of  $L$  and  $\Theta_{YX}$  respectively, we note that:

$$\begin{aligned}\sigma'_Y \geq \sigma_Y - \|L_Y - L_Y^*\|_2 &\geq \frac{C_1^2\lambda_n}{\omega_Y} m\psi^2(12\beta+1) - C_1\lambda_n \\ &\geq \frac{C_1^2\lambda_n}{\omega_Y} 12m\psi^2\beta \geq 8C_1\lambda_n \geq 8\|L_Y - L_Y^*\|_2 \\ \sigma'_{YX} \geq \sigma_{YX} - \|\Theta_{YX} - \Theta_{YX}^*\|_2 &\geq \frac{C_1^2\lambda_n\gamma^2}{\omega_{YX}} m\psi^2\left(\frac{6\beta}{\gamma} + 12\beta\right) - C_1\lambda_n\gamma \\ &\geq \frac{6C_1^2\lambda_n\gamma^2 m\psi^2\beta}{\omega_{YX}\gamma} - C_1\gamma\lambda_n \geq 8C_1\lambda_n\gamma \geq 8\|\Theta_{YX} - \Theta_{YX}^*\|_2\end{aligned}$$

Once again appealing to Proposition 2.2 in (Chandrasekaran et al., 2012), we have:

$$\begin{aligned}\Phi_{\delta,\gamma}(C_{T'}) &\leq m\|\mathcal{P}_{T(L_Y)^\perp}(L_Y - L_Y^*)\|_2 + m\|\mathcal{P}_{T(\Theta_{YX})^\perp}(\Theta_{YX} - \Theta_{YX}^*)\|_2 \\ &\leq m\frac{\|L_Y - L_Y^*\|_2^2}{\sigma'_Y} + m\frac{\|\Theta_{YX} - \Theta_{YX}^*\|_2^2}{\sigma'_{YX}} \\ &\leq \frac{mC_1^2\lambda_n^2}{\frac{C_1^2\lambda_n m\psi^2(12\beta+1)}{\omega_Y} - C_1\lambda_n} + \frac{mC_1^2\lambda_n^2\gamma^2}{\frac{C_1^2\lambda_n\gamma^2 m\psi^2(\frac{6\beta}{\gamma} + 12\beta)}{\omega_{YX}} - C_1\lambda_n\gamma} \\ &\leq \frac{\lambda_n\omega_Y}{12\beta\psi^2} + \frac{\lambda_n\omega_{YX}}{12\beta\psi^2} \leq \frac{\lambda_n}{6\beta\psi^2}\end{aligned}$$

This leads to the result that  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger\mathbb{I}^*\mathcal{A}C_{T'}] \leq \frac{\lambda_n}{6\beta}$ .

5. Following the same reasoning as step 4, we conclude:

$$\begin{aligned}\Phi_{\delta,\gamma}(C_{T'}) &\leq m\frac{\|L_Y - L_Y^*\|_2^2}{\sigma'_Y} + m\frac{\|\Theta_{YX} - \Theta_{YX}^*\|_2^2}{\sigma'_{YX}} \\ &\leq \frac{mC_1^2\lambda_n^2}{\frac{C_1^2\lambda_n m\psi^2(\frac{2}{C_2\psi^2} + 1)}{\omega_Y} - C_1\lambda_n} + \frac{mC_1^2\lambda_n^2\gamma^2}{\frac{C_1^2\lambda_n\gamma^2 m\psi^2(\frac{6\beta}{\gamma} + \frac{2}{C_2\psi^2})}{\omega_{YX}} - C_1\lambda_n\gamma} \\ &\leq \frac{C_2\omega_Y}{2}\lambda_n + \frac{C_2\omega_{YX}}{2}\lambda_n \leq C_2\lambda_n\end{aligned}$$

6. This fact follows immediately since  $\|S_Y - S_Y^*\|_\infty \leq \delta C_1\lambda_n$  and the smallest nonzero magnitude entry of  $S_Y^*$  is greater than  $2\delta C_1\lambda_n$ .  $\square$

### 5.1.2 From Variety Constraints to Tangent-Space Constraints

Consider any optimal solution  $(\Theta^{\mathcal{M}}, S_Y^{\mathcal{M}}, L_Y^{\mathcal{M}})$  of (5.2). In Corollary 5.2, we concluded that the variables  $(\Theta_{YX}^{\mathcal{M}}, S_Y^{\mathcal{M}}, L_Y^{\mathcal{M}})$  are smooth points of their respective varieties. As a result, the rank constraints  $\text{rank}(L_Y) \leq \text{rank}(L_Y^*)$  and  $\text{rank}(\Theta_{YX}) \leq \text{rank}(\Theta_{YX}^*)$  can be “linearized” to  $L_Y \in T(L_Y^{\mathcal{M}})$  and  $\Theta_{YX} \in T(\Theta_{YX}^{\mathcal{M}})$  respectively. Since all the remaining constraints are convex, the optimum of this linearized program is also the optimum of (5.2). Moreover, we once more appeal to Corollary 5.2 to conclude that the constraints in  $\mathcal{M}$  along  $\mathcal{P}_{T(L_Y^*)^\perp}$  and  $\mathcal{P}_{T(\Theta_{YX}^*)^\perp}$  are strictly feasible at  $(\Theta^{\mathcal{M}}, S_Y^{\mathcal{M}}, L_Y^{\mathcal{M}})$ . As a result, these constraints are locally inactive and can be removed without changing the optimum. Finally, we claim that the constraint  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \Delta] \leq 5\lambda_n$  in (5.2) can also be removed in this “linearized” convex program. In particular, letting  $\mathbb{H}_{\mathcal{M}} \triangleq \Omega(S_Y^*) \times T(L_Y^{\mathcal{M}}) \times T(\Theta_{YX}^{\mathcal{M}}) \times \mathbb{S}^q$  (note  $\mathbb{H}_{\mathcal{M}} \in U(\omega_Y, \omega_{YX})$  where the set  $U(\omega_Y, \omega_{YX})$  is defined in (2.3)), consider the following convex optimization program with the constraint  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \Delta] \leq 5\lambda_n$  removed :

$$\begin{aligned} (\tilde{\Theta}, \tilde{S}_Y, \tilde{L}_Y) = & \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \Theta \succ 0 \\ S_Y, L_Y \in \mathbb{S}^p}}{\text{argmin}} & -\ell(\Theta; \{X^{(i)}, Y^{(i)}\}_{i=1}^n) + \lambda_n[\delta\|S_Y\|_{\ell_1} + \|L_Y\|_* + \gamma\|\Theta_{YX}\|_*] \\ \text{s.t.} & \Theta_Y = S_Y - L_Y, \quad (S_Y, L_Y, \Theta_{YX}, \Theta_X) \in \mathbb{H}_{\mathcal{M}} \end{aligned} \quad (5.3)$$

In the following theorem, we prove that under additional conditions on the regularization parameter  $\lambda_n$  of (5.3), the set of variables  $(\Theta^{\mathcal{M}}, S_Y^{\mathcal{M}}, L_Y^{\mathcal{M}})$  is the unique optimum of (5.3).

**Proposition 5.3.** *Let  $C' = (2 + \delta \deg(S_Y^*) + \gamma)\psi$ ,  $C_1 = \frac{12}{\alpha} + \frac{1}{\psi^2}$ ,  $C_2 = \frac{4}{\alpha}(1 + \frac{1}{3\beta})$ , and  $C'_{\text{samp}} = \max\left\{\frac{1}{48\beta\psi}, 4C_2C', \frac{32m\psi C'^2 C_2}{\alpha}, 12\beta m\psi C'^2 C_1^2\right\}$ . Suppose that the number of observed samples obeys  $n \geq 4608\beta^2 m^2 \psi^2 C'^2_{\text{samp}}(p+q)$ , and the regularization parameter  $\lambda_n$  is chosen in the following range:  $\lambda_n \in \left[6\beta\sqrt{\frac{128(p+q)m^2\psi^2}{n}}, \frac{1}{C'_{\text{samp}}}\right]$ . Then, with probability greater than  $1 - 2\exp\left\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\right\}$ ,  $(\tilde{\Theta}, \tilde{S}_Y, \tilde{L}_Y) = (\Theta^{\mathcal{M}}, S_Y^{\mathcal{M}}, L_Y^{\mathcal{M}})$ .*

*Proof.* The high-level proof strategy is to show that the constraint  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \Delta] \leq 5\lambda_n$  is inactive at the optimum of (5.3). That is, we show that  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(\tilde{S}_Y - S_Y^*, \tilde{L}_Y - L_Y^*, \tilde{\Theta}_{YX} - \Theta_{YX}^*, \tilde{\Theta}_X - \Theta_X^*)] < 5\lambda_n$ . The proof of this fact relies on the results of the following lemmas:

**Lemma 5.4.** *Let  $\Delta = (\tilde{S}_Y - S_Y^*, \tilde{L}_Y - L_Y^*, \tilde{\Theta}_{YX} - \Theta_{YX}^*, \tilde{\Theta}_X - \Theta_X^*)$ . Denote  $R_{\Sigma^*}(\Delta) \triangleq \Sigma^* \left[ \sum_{k=2}^{\infty} (-\mathcal{A}(\Delta) \Sigma^{*-1})^k \right]$ . If  $\Phi_{\delta,\gamma}[\Delta] \leq \frac{1}{2C'}$ , then:  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger R_{\Sigma^*}(\Delta)] \leq 2m\psi C'^2 \Phi_{\delta,\gamma}[\Delta]^2$ .*

*Proof.* We begin by introducing a quantity that plays an important role in the proof of this lemma and was also employed in Chandrasekaran et al. (2012). Given a symmetric  $p \times p$  matrix  $M$ , we define the quantity  $\mu(\Omega(M))$  with respect to the tangent space  $\Omega(M)$  of the variety of  $p \times p$  matrices with at most  $|\text{support}(M)|$  nonzero entries:

$$\mu(\Omega(M)) \triangleq \max_{N \in \Omega(M), \|N\|_{\ell_\infty} = 1} \|N\|_2 \quad (5.4)$$

One can show that a sparse matrix  $M$  with “bounded degree” (a small number of non zeros per row/column) has small  $\mu(M)$ . Specifically, for any  $p \times p$  matrix  $M$ , we have  $\mu(\Omega(M)) \leq \deg(M)$  where  $\deg(M)$  is equal to the maximum number of non zeros in any column/row of  $M$ .



(Chandrasekaran et al., 2012). We now proceed with the proof:

$$\begin{aligned}
\Phi_{\delta,\gamma}[\mathcal{A}^\dagger R_{\Sigma^\star}(\Delta)] &\leq m\psi \left[ \sum_{k=2}^{\infty} (\psi \|\Delta\|_2)^k \right] \leq m\psi \sum_{k=2}^{\infty} \psi^k \left( \delta\mu(\Omega(S_Y^\star)) \frac{\|\Delta S_Y\|_{\ell_\infty}}{\delta} + \|\Delta L_Y\|_2 \right. \\
&\quad \left. + \gamma \frac{\|\Delta \Theta_{YX}\|_2}{\gamma} + \|\Delta \Theta_X\|_2 \right)^k \\
&\leq m\psi \sum_{k=2}^{\infty} \psi^k \left( \delta \deg(S_Y^\star) \frac{\|\Delta S_Y\|_{\ell_\infty}}{\delta} + \|\Delta L_Y\|_2 \right. \\
&\quad \left. + \gamma \frac{\|\Delta \Theta_{YX}\|_2}{\gamma} + \|\Delta \Theta_X\|_2 \right)^k \\
&\leq m\psi^3 \frac{(2 + \delta \deg(S_Y^\star) + \gamma)^2 \Phi_{\delta,\gamma}[\Delta]^2}{1 - (2 + \delta \deg(S_Y^\star) + \gamma) \Phi_{\delta,\gamma}[\Delta] \psi} \leq 2m\psi C'^2 \Phi_{\delta,\gamma}[\Delta]^2
\end{aligned}$$

Note that the second inequality employs the property that  $(\tilde{S}_Y - S_Y^\star) \in \Omega(S_Y^\star)$  and the quantity defined in (5.4).  $\square$

**Lemma 5.5.** *Let  $C_{T_{\mathcal{M}}} = \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}^\perp(0, L_Y^\star, \Theta_{YX}^\star, 0)$ . Furthermore, let  $E_n = \Sigma_n - \Sigma^\star$ , and  $\Delta = (\tilde{S}_Y - S_Y^\star, \tilde{L}_Y - L_Y^\star, \tilde{\Theta}_{YX} - \Theta_{YX}^\star, \tilde{\Theta}_X - \Theta_X^\star)$ . Finally, define:*

$$r = \max \left\{ \frac{4}{\alpha} \left( \Phi_{\delta,\gamma}[\mathcal{A}^\dagger E_n + \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} C_{T_{\mathcal{M}}}] + \lambda_n \right), \Phi_{\delta,\gamma}[C_{T_{\mathcal{M}}}] \right\} \quad (5.5)$$

If  $r \leq \min\{\frac{1}{4C'}, \frac{\alpha}{32m\psi C'^2}\}$ , then  $\Phi_{\delta,\gamma}[\Delta] \leq 2r$ .

*Proof.* The proof of this result uses Brouwer's fixed-point theorem, and is inspired by the proof of a similar result in (Ravikumar et al., 2008; Chandrasekaran et al., 2012). The optimality conditions of (5.3) suggest that there exist Lagrange multipliers  $Q_\Omega \in \Omega(S_Y^\star)^\perp$ ,  $Q_{T_Y} \in T(\tilde{L}_Y)^\perp$ , and  $Q_{T_{YX}} \in T(\tilde{\Theta}_{YX})^\perp$  such that

$$\begin{aligned}
[\Sigma_n - \tilde{\Theta}^{-1}]_Y + Q_\Omega &\in -\lambda_n \delta \partial \|\tilde{S}_Y\|_{\ell_1}; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_Y + Q_{T_Y} \in -\lambda_n \partial \|\tilde{L}_Y\|_\star \\
[\Sigma_n - \tilde{\Theta}^{-1}]_{YX} + Q_{T_{YX}} &\in -\lambda_n \gamma \partial \|\tilde{\Theta}_{YX}\|_\star; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_X = 0
\end{aligned}$$

Letting the SVD decomposition of  $\tilde{L}$  and  $\tilde{\Theta}_{YX}$  be given by  $\tilde{L}_Y = \bar{U} \bar{D} \bar{V}'$  and  $\tilde{\Theta}_{YX} = \check{U} \check{D} \check{V}'$  respectively, we can restrict the optimality conditions to the space  $\mathbb{H}_{\mathcal{M}}$  to obtain:

$$\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger (\Sigma_n - \tilde{\Theta}^{-1}) = (-\lambda_n \delta \text{sign}(\tilde{S}_Y), \lambda_n \bar{U} \bar{V}', -\lambda_n \gamma \check{U} \check{V}', 0) \quad (5.6)$$

Based on the Fisher information Assumption (2.5), the optimum of (5.3) is unique (this is because the Hessian of the negative log-likelihood term is positive definite restricted to the tangent space constraints). Moreover, using standard Lagrangian duality, one can show that the set of variables  $(\tilde{\Theta}, \tilde{S}_Y, \tilde{L}_Y)$  that satisfy (5.6) are unique. The matrix inversion lemma allows one to express  $\tilde{\Theta}^{-1}$  equivalently by:

$$\tilde{\Theta}^{-1} = [\Theta^\star + \mathcal{A}(\Delta)]^{-1} = \Sigma^\star - R_{\Sigma^\star}(\Delta) + \mathbb{I}^\star \mathcal{A}(\Delta)$$

Setting  $Z \triangleq (-\lambda_n \delta \text{sign}(\tilde{S}), \lambda_n \bar{U} \bar{V}', -\lambda_n \gamma \check{U} \check{V}', 0)$ , relation (5.6) can be restated as:

$$\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger (E_n - R_{\Sigma^\star}(\Delta) + \mathbb{I}^\star \mathcal{A}(\Delta)) = Z \quad (5.7)$$

Notice that  $\Phi_{\delta,\gamma}(Z) = \lambda_n$ . We now appeal to Brouwer's fixed-point theorem to bound  $\Phi_{\delta,\gamma}[\Delta]$ . Consider the following function  $G(\underline{\delta})$  with  $\underline{\delta} \in U(\omega_Y, \omega_{YX})$ :

$$G(\underline{\delta}) = \underline{\delta} - (\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}})^{-1} \left( \mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger [E_n - R_{\Sigma^*} \mathcal{A}(\underline{\delta} + C_{T_{\mathcal{M}}}) + \mathbb{I}^* \mathcal{A}(\underline{\delta} + C_{T_{\mathcal{M}}})] - Z \right)$$

Note that the function  $G(\underline{\delta})$  is well-defined since the operator  $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}$  is bijective due to Fisher information Assumption 1 in (2.5). As a result,  $\underline{\delta}$  is a fixed point of  $G(\underline{\delta})$  if and only if  $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger [E_n - R_{\Sigma^*} \mathcal{A}(\underline{\delta} + C_{T_{\mathcal{M}}}) + \mathbb{I}^* \mathcal{A}(\underline{\delta} + C_{T_{\mathcal{M}}})] = Z$ . Since the variables  $(\tilde{\Theta}, \tilde{S}_Y, \tilde{L}_Y)$  are the unique solution to (5.3), the only fixed point of  $G$  is  $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}[\Delta]$ . Next we show that this unique optimum lives inside the ball  $\mathbb{B}_r = \{\underline{\delta} \mid \Phi_{\delta,\gamma}(\underline{\delta}) \leq r, \underline{\delta} \in \mathbb{H}_{\mathcal{M}}\}$ . In particular, if we show that under the map  $G$ , the image of  $\mathbb{B}_r$  lies in  $\mathbb{B}_r$ , we can appeal to Brouwer's fixed point theorem to conclude that  $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}[\Delta] \in \mathbb{B}_r$ . For  $\underline{\delta} \in \mathbb{B}_r$ ,  $\Phi_{\delta,\gamma}[G(\underline{\delta})]$  can be bounded as follows:

$$\begin{aligned} \Phi_{\delta,\gamma}[G(\underline{\delta})] &= \Phi_{\delta,\gamma} \left[ \underline{\delta} - (\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}})^{-1} \left( \mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger [E_n - R_{\Sigma^*} \mathcal{A}(\underline{\delta} + C_{T_{\mathcal{M}}}) + \mathbb{I}^* \mathcal{A}(\underline{\delta} + C_{T_{\mathcal{M}}})] - Z \right) \right] \\ &= \Phi_{\delta,\gamma} \left[ (\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}})^{-1} \left( \mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger [E_n - R_{\Sigma^*} \mathcal{A}(\underline{\delta} + C_{T_{\mathcal{M}}}) + \mathbb{I}^* \mathcal{A} C_{T_{\mathcal{M}}}] - Z \right) \right] \\ &\leq \frac{1}{\alpha} \Phi_{\delta,\gamma} \left[ \mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger (E_n - R_{\Sigma^*} \mathcal{A}(\underline{\delta} + C_{T_{\mathcal{M}}}) + \mathbb{I}^* \mathcal{A}(C_{T_{\mathcal{M}}})) - Z \right] \\ &\leq \frac{2}{\alpha} \left[ \Phi_{\delta,\gamma}[\mathcal{A}^\dagger (E_n + \mathbb{I}^* \mathcal{A}(C_{T_{\mathcal{M}}}))] + \lambda_n \right] + \frac{2}{\alpha} \Phi_{\delta,\gamma}[\mathcal{A}^\dagger R_{\Sigma^*}(\underline{\delta})] \\ &\leq \frac{r}{2} + \frac{2}{\alpha} \Phi_{\delta,\gamma}[\mathcal{A}^\dagger R_{\Sigma^*}(\underline{\delta})] \end{aligned}$$

The first inequality holds because of Fisher information Assumption 1 in (2.5). The second inequality uses the property  $\Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\cdot)] \leq 2\Phi_{\delta,\gamma}(\cdot)$  and  $\Phi_{\delta,\gamma}(Z) = \lambda_n$ . Moreover, since  $r \leq \frac{1}{4C'}$ , we have  $\Phi_{\delta,\gamma}(\underline{\delta} + C_{T_{\mathcal{M}}}) \leq \Phi_{\delta,\gamma}(\underline{\delta}) + \Phi_{\delta,\gamma}(C_{T_{\mathcal{M}}}) \leq 2r \leq \frac{1}{2C'}$ . We can now appeal to Proposition 1 to obtain:

$$\begin{aligned} \frac{2}{\alpha} \Phi_{\delta,\gamma}[\mathcal{A}^\dagger R_{\Sigma^*}(\underline{\delta} + C_{T_{\mathcal{M}}})] &\leq \frac{4}{\alpha} m\psi C'^2 [\Phi_{\delta,\gamma}(\underline{\delta} + C_{T_{\mathcal{M}}})]^2 \\ &\leq \frac{16m\psi}{\alpha} C'^2 r^2 = \left[ \frac{16m\psi}{\alpha} C'^2 r \right] r \leq \frac{r}{2} \end{aligned}$$

Thus, we conclude that  $\Phi_{\delta,\gamma}[G(\underline{\delta})] \leq r$  and by Brouwer's fixed-point theorem,  $\Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] \leq r$ . Furthermore,

$$\Phi_{\delta,\gamma}[\Delta] \leq \Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] + \Phi_{\delta,\gamma}(C_{T_{\mathcal{M}}}) \leq 2r$$

□

**Lemma 5.6.** Suppose that the number of observed samples obeys  $n \geq 4608\beta^2 m^2 \psi^2 C_{\text{samp}}'^2 (p+q)$ , and the regularization parameter  $\lambda_n$  is chosen in the following range:  $\lambda_n \in \left[ 6\beta \sqrt{\frac{128(p+q)m^2 \psi^2}{n}}, \frac{1}{C_{\text{samp}}'} \right]$ . Then, with probability greater than  $1 - 2\exp\left\{ -\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2} \right\}$ ,  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{6\beta}$ .

*Proof.* First, note that  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger E_n] \leq m\|\Sigma_n - \Sigma^*\|_2$ . Using the results in (Davidson & Szarek, 2001) and the fact that  $\frac{\lambda_n}{6\beta} \leq 8\psi$  and  $n \geq \frac{2304(p+q)m^2 \psi^2}{\lambda_n^2}$ , the following bound holds:  $\Pr[m\|\Sigma_n - \Sigma^*\|_2 \geq \frac{\lambda_n}{6\beta}] \leq 2\exp\left\{ -\frac{n\lambda_n^2}{4608m^2 \psi^2} \right\}$ . Thus,  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{6\beta}$  with probability greater than  $1 - 2\exp\left\{ -\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2} \right\}$ . □

**Proof of Proposition 2:** We now proceed with completing the proof of this theorem. In particular, we show that  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(\tilde{S}_Y - S_Y^*, \tilde{L}_Y - L_Y^*, \tilde{\Theta}_{YX} - \Theta_{YX}^*, \tilde{\Theta}_X - \Theta_X^*)] < 5\lambda_n$ . Based on the optimality condition (5.7) and the property that  $\Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\cdot)] \leq 2\Phi_{\delta,\gamma}(\cdot)$ , we have:

$$\begin{aligned} \Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] &\leq 2\lambda_n + \Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger R_{\Sigma^*}(\Delta)] + \Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} C_{T_{\mathcal{M}}}] \\ &\quad + \Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger E_n] \\ &\leq 2\lambda_n + 2\Phi_{\delta,\gamma}[\mathcal{A}^\dagger R_{\Sigma^*}(\Delta)] + 2\Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} C_{T_{\mathcal{M}}}] \\ &\quad + 2\Phi_{\delta,\gamma}[\mathcal{A}^\dagger E_n] \end{aligned} \tag{5.8}$$

Appealing to Corollary 5.2 and Proposition 5.6, we have that  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} C_{T_{\mathcal{M}}}] \leq \frac{\lambda_n}{6\beta}$ ,  $\Phi_{\delta,\gamma}[C_{T'}] \leq C_2\lambda_n$  and (with high probability)  $\Phi_{\delta,\gamma}[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{6\beta}$ . Consequently, based on the bound on  $\lambda_n$  in assumption of Theorem 5.3, it is straightforward to show that  $r \leq \min\{\frac{1}{4C'}, \frac{\alpha}{32m\psi C'^2}\}$ . Hence by Proposition 5.5,  $\Phi_{\delta,\gamma}[\Delta] \leq \frac{1}{2C'}$ . Finally, we can appeal to Proposition 5.4 to obtain:

$$\Phi_{\delta,\gamma}[\mathcal{A}^\dagger R_{\Sigma^*}(\Delta)] \leq 2m\psi C'^2 \Phi_{\delta,\gamma}[\Delta]^2 \leq 2m\psi C'^2 C_1^2 \lambda_n^2 \leq \left[12\beta m\psi C'^2 C_1^2 \lambda_n\right] \frac{\lambda_n}{6\beta} \leq \frac{\lambda_n}{6\beta}$$

where the second to last bound comes from the bound on  $\lambda_n$ . Thus, the expression in (5.8) can be further simplified to:

$$\Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] \leq 2\lambda_n + 2\lambda_n \left( \frac{1}{6\beta} + \frac{1}{6\beta} + \frac{1}{6\beta} \right) \leq 2\lambda_n + \frac{\lambda_n}{\beta} \leq \frac{17\lambda_n}{8}$$

The last bound follows since  $\beta \geq 8$ . Furthermore,

$$\begin{aligned} \Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(\Delta)] &\leq \Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] + \Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^\perp} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] \\ &\quad + \Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}^\perp}(\Delta)] \\ &\leq \Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] + (1-\nu)\Phi_{\delta,\gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] \\ &\quad + \Phi_{\delta,\gamma}[\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A} C_{T_{\mathcal{M}}}] \\ &\leq \frac{17\lambda_n}{8} + \frac{17\lambda_n}{8}(1-\nu) + \frac{\lambda_n}{6\beta} < \frac{17\lambda_n}{8} + \frac{17\lambda_n}{8} + \frac{\lambda_n}{48} < 5\lambda_n \end{aligned}$$

Note that we appeal to Fisher information Assumption 2 in (2.6) in the second inequality.  $\square$

### 5.1.3 From Tangent Constraints to the Original Problem

Finally, we show that the tangent-space constraints in (5.3) can be removed without altering the optimum value. More concretely,

**Proposition 5.7.** *Suppose that the number of observed samples obeys  $n \geq 4608\beta^2 m^2 \psi^2 C_{smp}'^2 (p+q)$ , and the regularization parameter  $\lambda_n$  is chosen in the following range:*

*$\lambda_n \in \left[6\beta \sqrt{\frac{128(p+q)m^2\psi^2}{n}}, \frac{1}{C_{smp}'}\right]$ . Then, with probability greater than  $1 - 2\exp\left\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\right\}$ , the optimum of (5.3) is the same as the optimum of (5.1).*

*Proof.* Note that the bound on  $n$  implies that the assumptions of Proposition 5.3 are satisfied. Hence, we conclude that the solution of the tangent-space constrained program (5.3) is the same as the global optimum of the variety constrained program (5.2). Next, we show that the optimal variables of (5.3) remain unchanged once the tangent-space constraints are removed. We proceed by proving that the optimum set of variables  $(\tilde{\Theta}, \tilde{S}_Y, \tilde{L}_Y)$  satisfy the optimality conditions of (5.1) given by:

$$\begin{aligned} [\Sigma_n - \tilde{\Theta}^{-1}]_Y &\in -\lambda_n \delta \partial \|\tilde{S}_Y\|_{\ell_1}, \quad [\Sigma_n - \tilde{\Theta}^{-1}]_Y \in \lambda_n \partial \|\tilde{L}_Y\|_{\star} \\ [\Sigma_n - \tilde{\Theta}^{-1}]_{YX} &\in -\lambda_n \partial \|\tilde{\Theta}_{YX}\|_{\star}, \quad [\Sigma_n - \tilde{\Theta}^{-1}]_X = 0 \end{aligned}$$

Equivalently, we show that  $(\tilde{\Theta}, \tilde{S}_Y, \tilde{L}_Y)$  satisfy the following set of equations:

1.  $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger (\Sigma_n - \tilde{\Theta}^{-1}) = (-\lambda_n \gamma_1 \text{sign}(\tilde{S}_Y), \lambda_n UV' - \lambda_n \gamma_2 \check{U} \check{V}', 0)$
2.  $\Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger (\Sigma_n - \tilde{\Theta}^{-1})] < \lambda_n$

Here,  $UDV'$  is the SVD decomposition of  $\tilde{L}_Y$  and  $\check{U} \check{D} \check{V}'$  is the SVD decomposition of  $\tilde{\Theta}_{YX}$ . It is clear that the first condition is satisfied since the variables  $(\tilde{\Theta}, \tilde{S}_Y, \tilde{L}_Y)$  are optimal with respect to (5.3). To prove that the second condition is met, it suffices to show that:

$$\begin{aligned} \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] &< \lambda_n - \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger E_n] \\ &- \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger R_{\Sigma^\star}(\Delta)] - \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} C_{T_{\mathcal{M}}}] \end{aligned} \quad (5.9)$$

We first note:

$$\begin{aligned} \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] &\leq \lambda_n + 2\Phi_{\delta, \gamma}[\mathcal{A}^\dagger R_{\Sigma^\star}(\Delta)] + 2\Phi_{\delta, \gamma}[\mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} C_{T_{\mathcal{M}}}] \\ &+ 2\Phi_{\delta, \gamma}[\mathcal{A}^\dagger E_n] \\ &\leq \lambda_n + \frac{\lambda_n}{\beta} = \frac{(\beta + 1)\lambda_n}{\beta} \end{aligned}$$

Appealing to the Fisher information Assumption 1 in (2.5), we obtain:

$$\begin{aligned} \Phi_{\delta, \delta}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] &\leq \frac{(\beta + 1)\lambda_n}{\beta} \left(1 - \frac{3}{\beta + 1}\right) = \lambda_n - \frac{2\lambda_n}{\beta} < \lambda_n - \frac{\lambda_n}{2\beta} \\ &\leq \lambda_n - \Phi_{\delta, \gamma}[\mathcal{A}^\dagger R_{\Sigma}(\Delta)] - \Phi_{\delta, \gamma}[\mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} C_{T_{\mathcal{M}}}] - \Phi_{\delta, \gamma}[\mathcal{A}^\dagger E_n] \\ &\leq \lambda_n - \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger R_{\Sigma^\star}(\Delta)] - \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} C_{T_{\mathcal{M}}}] \\ &- \Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{A}^\dagger E_n] \end{aligned}$$

Here the last inequality holds since  $\Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\cdot)] \leq \Phi_{\delta, \gamma}(\cdot)$  □

**Proof of Theorem 1:** Letting  $\bar{m} = \max\{\delta, 1, \gamma\}$ , we note that  $C'$  in Proposition 5.3 can be bounded as follows:  $C' \leq 2\psi \deg(S_Y^\star) \bar{m}$ . Further, one can check that the constants in Proposition 5.3 are related to the constants in Theorem 2.1 as follows:  $C'_{\sigma_{YX}} \leq m C_{\sigma_{YX}}$ , and subsequently  $C'_{\text{samp}} \leq m \bar{m}^2 \deg(S_Y^\star) C_{\text{samp}}$ . Under the assumptions of Theorem 2.1, we can appeal to Corollary 5.2, Proposition 5.3, and Proposition 5.7 to conclude that with probability greater than  $1 - 2\exp\left\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\right\}$ , the variables  $(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y)$  are structurally correct estimates of  $(\Theta^\star, S_Y^\star, L_Y^\star)$ ; that is:  $\text{sign}(\hat{S}_Y) = \text{sign}(\hat{S}_Y^\star)$ ,  $\text{rank}(\hat{L}_Y) = \text{rank}(L_Y^\star)$ ,  $\text{rank}(\hat{\Theta}_{YX}) = \text{rank}(\Theta_{YX}^\star)$  and the estimates  $(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y)$  are “close” to population quantities in appropriate norms:  $\Phi_{\delta, \gamma}[\hat{S}_Y - S_Y^\star, \hat{L}_Y - L_Y^\star, \hat{\Theta}_{YX} - \Theta_{YX}^\star, \hat{\Theta}_X - \Theta_X^\star] \leq 2C_1 \lambda_n$ .

## 5.2 Choices of Parameters $\gamma, \delta$ in the estimator (1.6)

In this section, we give conditions on the population Fisher information  $\mathbb{I}^*$  such that Assumptions 1 and 2 in (2.5) and (2.6) are satisfied by a non-empty set of values of  $\delta, \gamma$ . In the subsequent discussion in this section, we employ the following notation to denote restrictions of a subspace  $\mathbb{H} = H_1 \times H_2 \times H_3 \times H_4 \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$  (here  $H_1, H_2, H_3, H_4$  are subspaces in  $\mathbb{S}^p, \mathbb{S}^p, \mathbb{R}^{p \times q}, \mathbb{S}^q$ , respectively) to its individual components. The restriction to the first component of  $\mathbb{H}$  is given by  $\mathbb{H}[1] = H_1 \times \{0\} \times \{0\} \times \{0\} \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ . The restrictions  $\mathbb{H}[2], \mathbb{H}[3], \mathbb{H}[4]$  to the other components of  $\mathbb{H}$  are defined in an analogous manner.

As our first quantity, we consider the minimum gain of  $\mathbb{I}^*$  restricted to each of the tangent spaces  $\Omega(S_Y^*)$ , and  $T_Y^*, T_{YX}^*$  separately:

$$\eta_1(\mathbb{H}^*; \omega_Y, \omega_{YX}) = \min_{\substack{\mathbb{H} \in U(\omega_Y, \omega_{YX}) \\ i=1,2,3,4}} \min_{\substack{M \in \mathbb{H}[i] \\ \|M\|_{\Phi_{1,1}}=1}} \|\mathcal{P}_{\mathbb{H}[i]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}_{\mathbb{H}[i]}(M)\|_{\Phi_{1,1}}. \quad (5.10)$$

Here the set  $U(\omega_Y, \omega_{YX})$  is defined in (2.3). Recall that this set denotes the distortions around the population tangent spaces  $T_Y^*, T_{YX}^*$ . Notice also that there is no appearance of  $\delta, \gamma$  in the norm  $\Phi$ . The quantity  $\eta_1(\mathbb{H}^*; \omega_Y, \omega_{YX})$  being large ensures that  $\mathbb{I}^*$  is well-conditioned when restricted to each of the tangent spaces  $\Omega(S_Y^*), T_Y', T_{YX}'$  separately. The second quantity we consider is the maximal inner-product between elements in each of the tangent spaces  $\Omega(S_Y^*), T_Y^*, T_{YX}^*$  and those in their respective orthogonal complements (again, in the metric induced by  $\mathbb{I}^*$ ):

$$\eta_2(\mathbb{H}^*; \omega_Y, \omega_{YX}) = \max_{\substack{\mathbb{H} \in U(\omega_Y, \omega_{YX}) \\ i=1,2,3,4}} \max_{\substack{M \in \mathcal{A}(\mathbb{H}[i])^\perp \\ \|M\|_2=1}} \|\mathcal{P}_{\mathcal{A}(\mathbb{H}[i])^\perp} \mathbb{I}^* \mathcal{P}_{\mathcal{A}(\mathbb{H}[i])}(M)\|_2 \quad (5.11)$$

One additional aspect of Assumptions 1 and 2 (in (2.5) and (2.6)) that is not addressed via the quantities  $\eta_1(\mathbb{H}^*; \omega_Y, \omega_{YX}), \eta_2(\mathbb{H}^*; \omega_Y, \omega_{YX})$  is the gain of the population Fisher information  $\mathbb{I}^*$  restricted to  $\Omega_Y^* \oplus T_Y^*$ . Controlling this gain ensures that the tangent spaces  $\Omega_Y^*$  and  $T_Y^*$  have a transverse intersection in the metric induced by  $\mathbb{I}^*$ ; as discussed in previous work by Chandrasekaran et al. (2012), such a property is critical to ensure the accurate estimation of the latent-variable graphical model specifying the conditional distribution of  $Y|f(X)$ . Following the approach adopted in that work, we control the gain of  $\mathbb{I}^*$  restricted to  $\Omega_Y^* \oplus T_Y^*$  via conditions involving three quantities. The first quantity  $\deg(S_Y^*)$  makes an appearance in Theorem 2.1 – it is the maximum number of nonzeros per row/column of  $S_Y^*$ , and it denotes the degree of the graphical model structure underlying the conditional distribution of  $Y|f(X), \zeta$ . The degree of the sparse component  $S_Y^*$  being small ensures that the graphical model underlying  $Y|f(X), \zeta$  is indeed a sparsely connected structure. Bounds on the degree of a population graphical model play an important role in the literature in results on consistent graphical model selection (Meinshausen & Bühlmann, 2006; Ravikumar et al., 2008). The second quantity is an *incoherence* parameter, which played an important role in the literature on low-rank matrix completion (Candès & Recht, 2009). Specifically, for a matrix  $N \in \mathbb{S}^{p_1}$ , the incoherence of the row-space / column-space of  $N$  is given by:

$$\text{inc}(N) \triangleq \max_{1 \leq i \leq p_1, 1 \leq j \leq p_1} \max \left\{ \|\mathcal{P}_{\text{column-space}(N)}(e_i)\|_{\ell_2}, \|\mathcal{P}_{\text{row-space}(N)}(e_j)\|_{\ell_2} \right\}, \quad (5.12)$$

where  $\mathcal{P}$  denotes the projection operation and  $e_i \in \mathbb{R}^{p_1}$  denotes the  $i$ 'th standard basis vector. The incoherence parameter of the low-rank matrix  $L_Y^*$  being small ensures that the latent variables  $\zeta$  affect most of the observed responses  $Y$ . As developed by Chandrasekaran et al. (2012), the quantities  $\deg(S_Y^*)$  and  $\text{inc}(L_Y^*)$  being small simultaneously ensures that the tangent spaces  $\Omega_Y^* \triangleq \Omega(S_Y^*)$  and  $T_Y^* \triangleq T(L_Y^*)$  are sufficiently transverse in the standard Euclidean inner-product. To

further ensure that the minimum gain of  $\mathbb{I}^*$  restricted to  $\Omega_Y^* \oplus T_Y^*$  is bounded below (i.e., to certify transversality of  $\Omega_Y^*$  and  $T_Y^*$  in the metric induced by  $\mathbb{I}^*$ ), Chandrasekaran et al. (2012) introduce the following quantity for  $\mathbb{W} = \mathbb{S}^p \times \{0\} \times \{0\} \times \{0\} \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ :

$$\eta_3(\Omega_Y^*, T_Y^*; \omega_Y) = \max \left\{ \max_{\rho(T_Y^*, T_Y^*) \leq \omega_Y} \max_{\substack{M \in T_Y^* \\ \|M\|_{\ell_\infty} = 1}} \|\mathcal{P}_{\mathcal{A}(\mathbb{W})} \mathbb{I}^* M\|_{\ell_\infty}, \max_{\substack{M \in \Omega_Y^* \\ \|M\|_2 = 1}} \|\mathcal{P}_{\mathcal{A}(\mathbb{W})} \mathbb{I}^* M\|_2 \right\} \quad (5.13)$$

The reason for the statement of this definition in terms of the  $\ell_\infty$  and spectral norms is that these are the dual norms of the regularizers employed in (1.6) (recall the discussion in Section 2.2). As shown by Chandrasekaran et al. (2012) and as described in the following proposition, suitably controlling the quantities  $\deg(S_Y^*)$ ,  $\text{inc}(L_Y^*)$ ,  $\eta_3(\Omega_Y^*, T_Y^*; \omega_Y)$  leads to lower bounds on the minimum gain of  $\mathbb{I}^*$  restricted to  $\Omega_Y^* \oplus T_Y^*$ , which enables the accurate estimation of the latent-variable graphical model underlying  $Y|f(X)$ .

In the following proposition, we describe a set of conditions on the quantities  $\eta_1(\mathbb{H}^*; [\omega_Y, \omega_{YX}])$ ,  $\eta_2(\mathbb{H}^*; [\omega_Y, \omega_{YX}])$ ,  $\eta_3(\Omega_Y^*, T_Y^*; \omega_Y)$ ,  $\deg(S_Y^*)$ , and  $\text{inc}(L_Y^*)$ , which lead to Assumptions 1 and 2 (14) and (15) (main theorem) being satisfied for  $(\delta, \gamma)$  inside a polyhedral set. We explicitly characterize this set and show that it is non-empty. For notational convenience, we denote  $\eta_1^* \triangleq \eta_1(\mathbb{H}^*; \omega_Y, \omega_{YX})$ ,  $\eta_2^* \triangleq \eta_2(\mathbb{H}^*; \omega_Y, \omega_{YX})$ , and  $\eta_3^* \triangleq \eta_3(\Omega_Y^*, T_Y^*; \omega_Y)$ .

**Proposition 5.8.** *Fix  $\alpha > 0, \nu \in (0, 1/3), \omega_Y > 0, \omega_{YX} > 0$ . Let  $\beta \triangleq \frac{3-\nu}{\nu}$ . Suppose that (i)  $\eta_1^* \geq 2\alpha$ , (ii)  $\eta_2^* \leq \min \left\{ \alpha(1 - \frac{3}{1+\beta}), \sqrt{\frac{\alpha}{\beta}} \frac{[2 \text{inc}(L_Y^*) + \omega_Y]}{4(1-\omega_Y)}, \frac{\alpha}{\beta\sqrt{2} \deg(S_Y^*)}, \frac{1}{2}(\frac{\alpha}{\beta})^{3/2} \right\}$ , (iii)  $\eta_3^* \leq \sqrt{\frac{\alpha}{\beta}}$ , and (iv)  $\frac{2 \text{inc}(L_Y^*) + \omega_Y}{1-\omega_Y} \deg(S_Y^*) \leq \frac{8\alpha}{\beta}$ . Then Assumptions 1 and 2 in (2.5) and (2.6) are satisfied for all  $(\delta, \gamma)$  in the following non-empty polyhedral set:*

$$V(\alpha, \nu, \omega_Y, \omega_{YX}) = \left\{ (\delta, \gamma) \mid \frac{[2 \text{inc}(L_Y^*) + \omega_Y]}{4(1-\omega_Y)} \sqrt{\frac{\beta}{\alpha}} \leq \delta \leq \frac{2}{\deg(S_Y^*)} \sqrt{\frac{\alpha}{\beta}}; \right. \\ \left. \max \left\{ 1, \eta_2^* \deg(S_Y^*) \delta \frac{2\beta}{\alpha} \right\} \leq \gamma \leq \frac{\min\{\delta, 1\} \alpha}{\eta_2^* \beta} \right\}.$$

Conditions analogous to (iii), (iv) appear in previous work on latent-variable graphical model selection (Chandrasekaran et al., 2012), and in our context they are useful for ensuring structurally correct estimates of the latent-variable graphical model corresponding to the conditional distribution of  $Y|f(X)$ . Conditions (i), (ii) are relevant for simultaneously obtaining structurally correct estimates of the smallest dimension reduction  $f(X)$  and of the latent-variable graphical model specifying  $Y|f(X)$  via the convex program (1.6). See Definition 2.1 for more details.

*Proof.* The proof of this proposition relies on two quantities. Given a matrix  $M \in \mathbb{S}^{p_1}$ , we defined the first quantity  $\mu(\Omega(M))$  in (5.4) with respect to the tangent space  $\Omega(M)$ . Additionally we define the quantity  $\xi(T(M))$  with respect to the tangent space  $T(M)$ :

$$\xi(T(M)) \triangleq \max_{N \in T(M), \|N\|_2 = 1} \|N\|_\infty.$$

For extensive discussion regarding the properties of these quantities, we refer the reader to (Chandrasekaran et al., 2012). Here, we highlight a few important facts. In particular, one can check that

$\xi(T(M)) \in [\text{inc}(M), 2\text{inc}(M)]$  and  $\mu(\Omega(M)) \leq \deg(M)$  (recall that  $\text{inc}(M)$  measures the incoherence of  $M$  and  $\deg(M)$  denotes the maximum number of nonzero elements in any column or row of  $M$ ). Furthermore, given two linear subspaces  $T_1$  and  $T_2$ , the quantity  $\rho(T_1, T_2)$  that measures the distortion of tangent spaces (see the definition in (2.3)) allows us to bound the variation in  $\xi(T_2)$  as follows (Chandrasekaran et al., 2012):

$$\xi(T_2) \leq \frac{1}{1 - \rho(T_1, T_2)} [\xi(T_1) + \rho(T_1, T_2)]$$

Returning to the proof, we show that for any  $(\delta, \gamma)$  inside the polyhedron set  $V(\alpha, \nu, \omega_Y, \omega_{YX})$ , Fisher Assumption 1 and 2 in (2.5) and (2.6) are satisfied. First, using conditions (ii) and (iv), one can check that the polyhedron set  $V(\alpha, \nu, \omega_Y, \omega_{YX})$  is non-empty. Now, let  $\mathbb{H} = \Omega_Y^* \times T_Y' \times T_{YX}' \times \mathbb{S}^q$  be any subspace inside  $U(\omega_Y, \omega_{YX})$  and the tradeoff parameters be chosen so that  $(\delta, \gamma) \in V(\alpha, \nu, \omega_Y, \omega_{YX})$ . Further let  $\mathbb{Z} = \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ , and let  $(S_Y, L_Y, \Theta_{YX}, \Theta_X) \in \mathbb{H}$  with  $\|S_Y\|_{\ell_\infty} \leq \delta$ ,  $\|L_Y\|_2 \leq 1$ ,  $\|\Theta_{YX}\|_2 \leq \gamma$ ,  $\|\Theta_X\|_2 \leq 1$ . Suppose equality holds in at least one of these set of inequalities so that  $\Phi_{\delta, \gamma}(S_Y, L_Y, \Theta_{YX}, \Theta_X) = 1$ . Then, at least one of the following cases is active:

1. Suppose  $\|S_Y\|_{\ell_\infty} = \delta$ . Then using conditions (i) – (iii) of Proposition 5.8, we have:

$$\begin{aligned} \frac{1}{\delta} \|\mathcal{P}_{\mathbb{H}[1]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(S_Y, L_Y, \Theta_{YX}, \Theta_X)\|_{\Phi_{1,1}} &\geq \frac{1}{\delta} \left[ \|\mathcal{P}_{\mathbb{H}[1]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(S_Y, 0, 0, 0)\|_{\Phi_{1,1}} \right. \\ &\quad - \|\mathcal{P}_{\mathcal{A}(\mathbb{H}[1])} \mathbb{I}^* \mathcal{A}(0, L_Y, 0, 0)\|_{\ell_\infty} \\ &\quad \left. - \|\mathcal{P}_{\mathcal{A}(\mathbb{H}[1])} \mathbb{I}^* \mathcal{A}(0, 0, \Theta_{YX}, \Theta_X)\|_{\ell_\infty} \right] \\ &\geq \frac{1}{\delta} \left[ \|\mathcal{P}_{\mathbb{H}[1]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(S_Y, 0, 0, 0)\|_{\Phi_{1,1}} \right. \\ &\quad - \|\mathcal{P}_{\mathcal{A}(\mathbb{Z}[1])} \mathbb{I}^* \mathcal{A}(0, L_Y, 0, 0)\|_{\ell_\infty} \\ &\quad \left. - \|\mathcal{P}_{\mathcal{A}(\mathbb{Z}[1])} \mathbb{I}^* \mathcal{A}(0, 0, \Theta_{YX}, \Theta_X)\|_{\ell_\infty} \right] \\ &\geq 2\alpha - \frac{\eta_3^* \xi(T_Y')}{\delta} - \frac{2\eta_2^* \max\{\gamma, 1\}}{\delta} \\ &\geq 2\alpha - \frac{(2 \text{inc}(\Omega_Y^*) + \omega_Y) \eta_3^*}{(1 - \omega_Y) \delta} - \frac{2 \max\{\gamma, 1\} \eta_2^*}{\delta} \\ &\geq 2\alpha - \frac{4\alpha}{\beta} - \frac{2\alpha}{\beta} \geq 2\alpha - \frac{8\alpha}{\beta} \end{aligned}$$

2. Suppose  $\|L_Y\|_2 = 1$ . Then using conditions (i) – (iii) of Proposition 5.8, we have:

$$\begin{aligned} \|\mathcal{P}_{\mathbb{H}[2]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(S_Y, L_Y, \Theta_{YX}, \Theta_X)\|_{\Phi_{1,1}} &\geq \|\mathcal{P}_{\mathbb{H}[2]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(0, L_Y, 0, 0)\|_{\Phi_{1,1}} \\ &\quad - \|\mathcal{P}_{\mathcal{A}(\mathbb{H}[2])} \mathbb{I}^* \mathcal{A}(S_Y, 0, 0, 0)\|_2 \\ &\quad - \|\mathcal{P}_{\mathcal{A}(\mathbb{H}[2])} \mathbb{I}^* \mathcal{A}(0, 0, \Theta_{YX}, \Theta_X)\|_2 \\ &\geq \|\mathcal{P}_{\mathbb{H}[2]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(0, L_Y, 0, 0)\|_{\Phi_{1,1}} \\ &\quad - 2\|\mathcal{P}_{\mathcal{A}(\mathbb{Z}[2])} \mathbb{I}^* \mathcal{A}(S_Y, 0, 0, 0)\|_2 \\ &\quad - 2\|\mathcal{P}_{\mathcal{A}(\mathbb{Z}[2])} \mathbb{I}^* \mathcal{A}(0, 0, \Theta_{YX}, \Theta_X)\|_2 \\ &\geq 2\alpha - 2\eta_3^* \mu(\Omega_Y^*) \delta - 4\eta_2^* \max\{\gamma, 1\} \\ &\geq 2\alpha - 2\eta_3^* \deg(\Omega_Y^*) \delta - 4\eta_2^* \max\{\gamma, 1\} \\ &\geq 2\alpha - \frac{4\alpha}{\beta} - \frac{4\alpha}{\beta} \geq 2\alpha - \frac{8\alpha}{\beta} \end{aligned}$$

3. Suppose  $\|\Theta_{YX}\|_2 = \gamma$ . Then using conditions (i) and (ii) of Proposition 5.8, we have:

$$\begin{aligned}
\frac{1}{\gamma} \|\mathcal{P}_{\mathbb{H}[3]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(S_Y, L_Y, \Theta_{YX}, \Theta_X)\|_{\Phi_{1,1}} &\geq \frac{1}{\gamma} \left[ \|\mathcal{P}_{\mathbb{H}[3]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(0, 0, \Theta_{YX}, 0)\|_{\Phi_{1,1}} \right. \\
&\quad \left. - \|\mathcal{P}_{\mathcal{A}(\mathbb{H}[3])} \mathbb{I}^* \mathcal{A}(S_Y, L_Y, 0, \Theta_{YX})\|_2 \right] \\
&\geq \frac{1}{\gamma} \left[ \|\mathcal{P}_{\mathbb{H}[3]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(0, 0, \Theta_{YX}, 0)\|_{\Phi_{1,1}} \right. \\
&\quad \left. - 2\|\mathcal{P}_{\mathcal{A}(\mathbb{Z}[3])} \mathbb{I}^* \mathcal{A}(S_Y, L_Y, 0, \Theta_{YX})\|_2 \right] \\
&\geq 2\alpha - \frac{2\eta_2^* \mu(\Omega_Y^*) \delta}{\gamma} - \frac{4\eta_2^*}{\gamma} \\
&\geq 2\alpha - \frac{2\eta_2^* \deg(S_Y^*) \delta}{\gamma} - 4\eta_2^* \\
&\geq 2\alpha - \frac{\alpha}{\beta} - \frac{\alpha}{\beta} \geq 2\alpha - \frac{8\alpha}{\beta}
\end{aligned}$$

4. Suppose  $\|\Theta_X\|_2 = 1$ . Then using conditions (i) and (ii) of Proposition 5.8, we have:

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{H}[4]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(S_Y, L_Y, \Theta_{YX}, \Theta_X)\|_{\Phi_{1,1}} &\geq \|\mathcal{P}_{\mathbb{H}[4]} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(0, 0, 0, \Theta_X)\|_{\Phi_{1,1}} \\
&\quad - \|\mathcal{P}_{\mathcal{A}(\mathbb{H}[4])} \mathbb{I}^* \mathcal{A}(S_Y, L_Y, \Theta_{YX}, 0)\|_2 \\
&\geq 2\alpha - \eta_2^* \mu(\Omega_Y^*) \delta - 2\eta_2^* \max\{\gamma, 1\} \\
&\geq 2\alpha - \eta_2^* \deg(S_Y^*) \delta - 2\eta_2^* \gamma \\
&\geq 2\alpha - \frac{4\alpha}{\beta} - \frac{2\alpha}{\beta} \geq 2\alpha - \frac{8\alpha}{\beta}
\end{aligned}$$

From these results, we conclude that  $\Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}} \mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(S_Y, L_Y, \Theta_{YX}, \Theta_X)] \geq 2\alpha - \frac{8\alpha}{\beta}$ . Further, we can bound the quantity  $\chi(\mathbb{H}, \|\cdot\|_{\Phi_{\delta, \gamma}})$  in (2.1) as follows

$$\chi(\mathbb{H}, \|\cdot\|_{\Phi_{\delta, \gamma}}) \geq 2\alpha - \frac{8\alpha}{\beta} \geq \alpha \quad (5.14)$$

Using a similar decoupling technique, one can show:

$$\Phi_{\delta, \gamma}[\mathcal{P}_{\mathbb{H}^\perp} [\mathcal{A}^\dagger \mathbb{I}^* \mathcal{A}(S_Y, L_Y, \Theta_{YX}, \Theta_X)]] \leq \eta_2^* + \frac{8\alpha}{\beta} \leq \alpha(1 - \frac{3}{1+\beta}) + \frac{8\alpha}{\beta}$$

Using this bound and the bound on  $\chi(\mathbb{H}, \|\cdot\|_{\Phi_{\delta, \gamma}})$ , we control the quantity  $\varphi(\mathbb{H}, \|\cdot\|_{\Phi_{\delta, \gamma}})$  in (2.2):

$$\varphi(\mathbb{H}, \|\cdot\|_{\Phi_{\delta, \gamma}}) \leq \frac{\left(1 - \frac{3}{1+\beta}\right)\alpha + \frac{8\alpha}{\beta}}{2\alpha - \frac{8\alpha}{\beta}} \leq 1 - \frac{3}{1+\beta} \quad (5.15)$$

Since the bounds (5.14) and (5.15) are valid for all  $\mathbb{H} \in U(\omega_Y, \omega_{YX})$ , Fisher information Assumptions 1 and 2 (in (2.5) and (2.6)) are satisfied for  $(\delta, \gamma)$  inside the polyhedron set  $V(\alpha, \nu, \omega_Y, \omega_{YX})$ .  $\square$



### 5.3 High-Dimensional Consistency of the Estimator (1.7)

In this section, we discuss the consistency properties of the estimator (1.7) in a high-dimensional scaling regime. Specifically, suppose we observe samples  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n \subset \mathbb{R}^{p+q}$  of a collection of jointly Gaussian responses and covariates  $(Y, X)$  with joint population precision matrix  $\Theta^* = \begin{pmatrix} S_Y^* - L_Y^* & \Theta_{YX}^* \\ \Theta_{YX}^{*'} & \Theta_X^* \end{pmatrix} \subset \mathbb{S}^{p+q}$ , where  $S_Y^*$  is sparse,  $L_Y^*$  is low-rank, and  $\Theta_{YX}^*$  is column-sparse. Supplying these observations into the program (1.7) and obtaining estimates  $(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y) \subset \mathbb{S}^{p+q} \times \mathbb{S}^p \times \mathbb{S}^p$ , we prove in Theorem 5.9 that (under certain conditions on  $\Theta^*$  and with high probability) (a) the column support of  $\hat{\Theta}_{YX}$  is equal to the column support of  $\Theta_{YX}^*$ , (b)  $\text{rank}(\hat{L}_Y) = \text{rank}(L_Y^*)$ , and (c)  $\text{sign}(\hat{S}_Y) = \text{sign}(S_Y^*)$ . Thus, the subset of covariates that are sufficient for predicting the responses and the latent-variable graphical model specifying the conditional distribution of the responses given the covariates are both correctly identified.

Proceeding in a similar manner as in Section 2.2, we prove that the estimator (1.7) is consistent under assumptions on the conditioning of the population Fisher information  $\mathbb{I}^*$ . These assumptions are stated in terms of tangent spaces of the algebraic variety of *column-sparse* matrices. Letting  $M \in \mathbb{R}^{p \times q}$  be a matrix with  $k$  nonzero columns, the tangent space at  $M$  with respect to the variety of  $p \times q$  matrices with at most  $k$  nonzero columns is given by:

$$F(M) \triangleq \{J \in \mathbb{R}^{p \times q} \mid \text{columnsupport}(J) \subseteq \text{columnsupport}(M)\}.$$

Here ‘columnsupport’ denotes the indices of the nonzero columns. As in Section 2.2, we control the conditioning of  $\mathbb{I}(\Theta^*)$  for all subspaces  $\mathbb{H}'$  in the following set:<sup>5</sup>

$$\tilde{U}(\omega_Y) \triangleq \left\{ \Omega(S_Y^*) \times T_Y' \times F(\Theta_{YX}^*) \times \mathbb{S}^q \mid \rho(T_Y', T(L_Y^*)) \leq \omega_Y \right\}. \quad (5.16)$$

We control the quantities  $\chi(\mathbb{H}', \tilde{\Phi}_{\delta, \gamma})$  and  $\varphi(\mathbb{H}', \tilde{\Phi}_{\delta, \gamma})$  (defined in (2.1) and (2.2)) for all  $\mathbb{H} \in \tilde{U}(\omega_Y)$  and for  $\tilde{\Phi}_{\delta, \gamma}(S_Y, L_Y, \Theta_{YX}, \Theta_X)$  defined as:

$$\tilde{\Phi}_{\delta, \gamma}(S_Y, L_Y, \Theta_{YX}, \Theta_X) \triangleq \max \left\{ \frac{\|S_Y\|_{\ell_\infty}}{\delta}, \|L_Y\|_2, \frac{\|\Theta_{YX}\|_{2, \infty}}{\gamma}, \|\Theta_X\|_2 \right\}. \quad (5.17)$$

As with  $\Phi_{\delta, \gamma}$  in Section 2.2, the norm  $\tilde{\Phi}_{\delta, \gamma}$  is a slight variant of the dual norm of the regularizer  $\delta\|S_Y\|_{\ell_1} + \text{trace}(L_Y) + \gamma\|\Theta_{YX}\|_{1, 2}$  in (1.7).

In summary, given  $(\delta, \gamma, \omega_Y) \in \mathbb{R}_+ \times \mathbb{R}_+ \times (0, 1)$  we assume that the population Fisher information  $\mathbb{I}^*$  satisfies the following conditions:

$$\begin{aligned} \text{Assumption 3} & : \quad \inf_{\mathbb{H}' \in \tilde{U}(\omega_Y)} \chi(\mathbb{H}', \tilde{\Phi}_{\delta, \gamma}) \geq \alpha, \quad \text{for some } \alpha > 0 \\ \text{Assumption 4} & : \quad \sup_{\mathbb{H}' \in \tilde{U}(\omega_Y)} \varphi(\mathbb{H}', \tilde{\Phi}_{\delta, \gamma}) \leq 1 - \nu, \quad \text{for some } \nu \in (0, 1/3). \end{aligned}$$

As with the notation preceding the statement of Theorem 2.1, let  $\tau_Y$  denote the minimum nonzero entry in magnitude of  $S_Y^*$ , let  $\sigma_Y$  denote the minimum nonzero singular value of  $L_Y^*$ , and let  $\deg(S_Y^*)$  denote the maximal number of nonzeros per row/column of  $S_Y^*$ . Further, let  $\zeta_{YX}$  denote the minimum  $\ell_2$  norm over nonzero columns of  $\Theta_{YX}^*$  and let  $\kappa$  be the number of nonzero columns of  $\Theta_{YX}^*$ .

<sup>5</sup>The variety of column-sparse matrices is locally flat around  $\Theta_{YX}^*$  so that the tangent spaces at all points in a neighborhood of  $\Theta_{YX}^*$  are all equal to  $F(\Theta_{YX}^*)$ .

**Theorem 5.9.** Suppose we are given i.i.d observations  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n \subset \mathbb{R}^{p+q}$  of a collection of jointly Gaussian covariates/responses with population precision matrix  $\Theta^* \in \mathbb{S}_{++}^{p+q}$ . Fix  $\alpha > 0, \nu \in (0, 1/3), \omega_Y \in (0, 1)$ . Suppose the trade-off parameters  $\delta$  and  $\gamma$  are chosen such that the population Fisher information  $\mathbb{I}(\Theta^*)$  satisfies Assumptions 3 and 4.

Let  $m \triangleq \max\{\frac{1}{\delta}, 1, \frac{1}{\gamma}\}$ ,  $\bar{m} \triangleq \max\{\delta, 1, \gamma\}$ ,  $\beta \triangleq \frac{3-\nu}{\nu}$ , and  $\psi \triangleq \|(\Theta^*)^{-1}\|_2$ . Further, let  $C_1 = \frac{24}{\alpha} + \frac{1}{\psi^2}$ ,  $C_2 = \frac{8}{\alpha}(\frac{1}{3\beta} + 1)$ ,  $C_\sigma = C_1^2 \psi^2 \max\{12\beta + 1, \frac{1}{C_2 \psi^2} + 1\}$ ,  $C_{smp} = \max\{\frac{1}{48\psi\beta}, 48\beta\psi^3 C_1^2, 8\psi C_2, \frac{128\psi^3 C_2}{\alpha}\}$ , and  $\lambda_{upper} = \frac{1}{m\bar{m}^2 \max\{\deg(S_Y^*), \kappa\} C_{smp}}$ . Suppose that the following conditions hold:

1.  $n \geq \frac{4608\psi^2\beta^2 m^2(p+q)}{\lambda_{upper}^2}$ ; that is  $n \gtrsim \left[\frac{\beta^4}{\alpha^2} m^4 \bar{m}^4 \max\{\deg(S_Y^*)^2, \kappa^2\}\right](p+q)$
2.  $\lambda_n \in \left[\sqrt{\frac{4608\psi^2\beta^2 m^2(p+q)}{n}}, \lambda_{upper}\right]$ ; e.g.  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$
3.  $\tau_Y \geq 2\gamma_1 C_1 \lambda_n$ ; that is  $\tau_Y \gtrsim \frac{\beta}{\alpha} m \bar{m} \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$
4.  $\sigma_Y \geq m C_\sigma \lambda_n$ ; that is  $\sigma_Y \gtrsim \frac{\beta^2}{\alpha \omega_Y} m^2 \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$
5.  $\zeta_{YX} \geq 2\gamma C_1 \lambda_n$ ; that is  $\zeta_{YX} \gtrsim \frac{\beta}{\alpha} m \bar{m} \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$

Then with probability greater than  $1 - 2\exp\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\}$ , the optimal solution  $(\hat{\Theta}, \hat{S}_Y, \hat{L}_Y)$  of (1.7) with the observations  $\{Y^{(i)}, X^{(i)}\}_{i=1}^n$  satisfies the following properties:

1.  $\text{sign}(\hat{S}_Y) = \text{sign}(S_Y^*)$ ,  $\text{rank}(\hat{L}_Y) = \text{rank}(L_Y^*)$ , and  $\text{columnsupport}(\hat{\Theta}_{YX}) = \text{columnsupport}(\Theta_{YX}^*)$ .
2.  $\Phi_{\delta, \gamma}(\hat{S}_Y - S_Y^*, \hat{L}_Y - L_Y^*, \hat{\Theta}_{YX} - \Theta_{YX}^*, \hat{\Theta}_X - \Theta_X^*) \leq C_1 \lambda_n$ ; that is  $\|\hat{S}_Y - S_Y^*\|_{\ell_\infty} \lesssim \frac{\beta}{\alpha} m \delta \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{L}_Y - L_Y^*\|_2 \lesssim \frac{\beta}{\alpha} m \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{\Theta}_{YX} - \Theta_{YX}^*\|_{2, \infty} \lesssim \frac{\beta}{\alpha} \gamma m \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{\Theta}_X - \Theta_X^*\|_2 \lesssim \frac{\beta}{\alpha} m \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$ .

The strategy for the proof of Theorem 5.9 is analogous to that of Theorem 2.1.